LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

NLPContributions: An Annotation Scheme for Machine Reading of Scholarly Contributions in Natural Language Processing Literature

Jennifer D'Souza and Sören Auer
Technische Informationsbibliothek (TIB)
Welfengarten 1B // 30167 Hannover

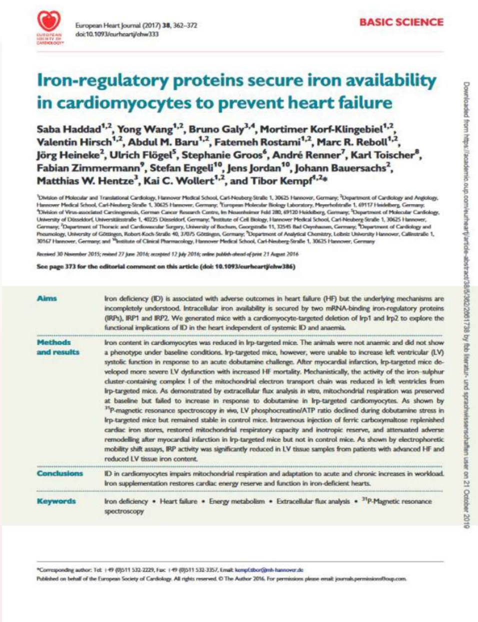
Leibniz Gemeinschaft

What if ...

- The global scientific knowledge base would be more than a document repository
- Scientific information and knowledge would be FAIR also for machines
  - The FAIR data principles are a set of guiding principles in order to make scientific data findable, accessible, interoperable, and reusable in the current digital ecosystem (Wilkinson et al. 2016)
- Currently
  - Findability could be better
  - Assuming OA, accessibility is OK
  - Interoperability and Reusability is non-existent
- The problem: The scholarly communications format is stuck in the last century
  - We have managed to digitize documents that used to be in print
  - While other areas have seen a transformative digitalization

Our Objective

- To foster the *digitalization* of digitized scholarly articles



European Heart Journal (2017) 38, 362–372
doi:10.1093/eurheartj/ehw333

BASIC SCIENCE

Iron-regulatory proteins secure iron availability in cardiomyocytes to prevent heart failure

Saba Haddad[1,2], Yong Wang[1,2], Bruno Galy[3,4], Mortimer Korf-Klingebiel[1,2], Valentin Hirsch[1,2], Abdul M. Baru[1,2], Fatemeh Rostami[1,2], Marc R. Reboll[1,2], Jörg Heineke[2], Ulrich Flögel[5], Stephanie Groos[6], André Renner[7], Karl Toischer[8], Fabian Zimmermann[9], Stefan Engeli[10], Jens Jordan[10], Johann Bauersachs[2], Matthias W. Hentze[3], Kai C. Wollert[1,2], and Tibor Kempf[1,2]*

[1]Division of Molecular and Translational Cardiology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; [2]Department of Cardiology and Angiology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; [3]European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany; [4]Division of Virus-associated Carcinogenesis, German Cancer Research Centre, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany; [5]Department of Molecular Cardiology, University of Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany; [6]Institute of Cell Biology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; [7]Department of Thoracic and Cardiovascular Surgery, University of Bochum, Georgstraße 11, 32545 Bad Oeynhausen, Germany; [8]Department of Cardiology and Pneumology, University of Göttingen, Robert-Koch-Straße 40, 37075 Göttingen, Germany; [9]Department of Analytical Chemistry, Leibniz University Hannover, Callinstraße 1, 30167 Hannover, Germany; and [10]Institute of Clinical Pharmacology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany

Received 30 November 2015; revised 27 June 2016; accepted 12 July 2016; online publish-ahead-of-print 21 August 2016

See page 373 for the editorial comment on this article (doi: 10.1093/eurheartj/ehw386)

Aims Iron deficiency (ID) is associated with adverse outcomes in heart failure (HF) but the underlying mechanisms are incompletely understood. Intracellular iron availability is secured by two mRNA-binding iron-regulatory proteins (IRPs), IRP1 and IRP2. We generated mice with a cardiomyocyte-targeted deletion of Irp1 and Irp2 to explore the functional implications of ID in the heart independent of systemic ID and anaemia.

Methods and results Iron content in cardiomyocytes was reduced in Irp-targeted mice. The animals were not anaemic and did not show a phenotype under baseline conditions. Irp-targeted mice, however, were unable to increase left ventricular (LV) systolic function in response to an acute dobutamine challenge. After myocardial infarction, Irp-targeted mice developed more severe LV dysfunction with increased HF mortality. Mechanistically, the activity of the iron-sulphur cluster-containing complex I of the mitochondrial electron transport chain was reduced in left ventricles from Irp-targeted mice. As demonstrated by extracellular flux analysis *in vitro*, mitochondrial respiration was preserved at baseline but failed to increase in response to dobutamine in Irp-targeted cardiomyocytes. As shown by ^{31}P-magnetic resonance spectroscopy in vivo, LV phosphocreatine/ATP ratio declined during dobutamine stress in Irp-targeted mice but remained stable in control mice. Intravenous injection of ferric carboxymaltose replenished cardiac iron stores, restored mitochondrial respiratory capacity and inotropic reserve, and attenuated adverse remodelling after myocardial infarction in Irp-targeted mice but not in control mice. As shown by electrophoretic mobility shift assays, IRP activity was significantly reduced in LV tissue samples from patients with advanced HF and reduced LV tissue iron content.

Conclusions ID in cardiomyocytes impairs mitochondrial respiration and adaptation to acute and chronic increases in workload. Iron supplementation restores cardiac energy reserve and function in iron-deficient hearts.

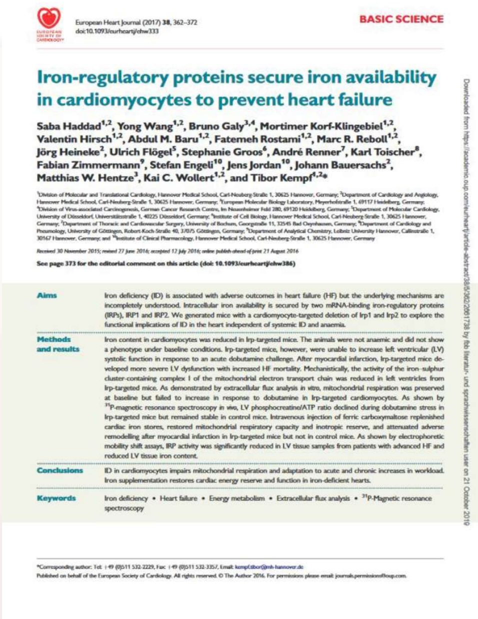
Keywords Iron deficiency • Heart failure • Energy metabolism • Extracellular flux analysis • ^{31}P-Magnetic resonance spectroscopy

*Corresponding author: Tel: +49 (0)511 532-2229, Fax: +49 (0)511 532-3357, Email: kempf.tibor@mh-hannover.de

Open Research Knowledge Graph

Our Objective

- To structure, in a fine-grained manner, knowledge elements from unstructured scholarly articles as a Knowledge Graph



European Heart Journal (2017) 38, 362–372
doi:10.1093/eurheartj/ehw333

BASIC SCIENCE

Iron-regulatory proteins secure iron availability in cardiomyocytes to prevent heart failure

Saba Haddad[1,2], Yong Wang[1,2], Bruno Galy[3,4], Mortimer Korf-Klingebiel[1,2], Valentin Hirsch[1,2], Abdul M. Baru[1,2], Fatemeh Rostami[1,2], Marc R. Reboll[1,2], Jörg Heineke[2], Ulrich Flögel[5], Stephanie Groos[6], André Renner[7], Karl Toischer[8], Fabian Zimmermann[9], Stefan Engeli[10], Jens Jordan[10], Johann Bauersachs[2], Matthias W. Hentze[3], Kai C. Wollert[1,2], and Tibor Kempf[1,2*]

[1]Division of Molecular and Translational Cardiology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; [2]Department of Cardiology and Angiology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; [3]European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany; [4]Division of Virus-associated Carcinogenesis, German Cancer Research Centre, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany; [5]Department of Molecular Cardiology, University of Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany; [6]Institute of Cell Biology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; [7]Department of Thoracic and Cardiovascular Surgery, University of Bochum, Georgstraße 11, 32545 Bad Oeynhausen, Germany; [8]Department of Cardiology and Pneumology, University of Göttingen, Robert-Koch Straße 40, 37075 Göttingen, Germany; [9]Department of Analytical Chemistry, Leibniz University Hannover, Callinstraße 1, 30167 Hannover, Germany; and [10]Institute of Clinical Pharmacology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany

Received 30 November 2015; revised 27 June 2016; accepted 12 July 2016; online publish-ahead-of-print 21 August 2016

See page 373 for the editorial comment on this article (doi: 10.1093/eurheartj/ehw386)

Aims Iron deficiency (ID) is associated with adverse outcomes in heart failure (HF) but the underlying mechanisms are incompletely understood. Intracellular iron availability is secured by two mRNA-binding iron-regulatory proteins (IRPs), IRP1 and IRP2. We generated mice with a cardiomyocyte-targeted deletion of Irp1 and Irp2 to explore the functional implications of ID in the heart independent of systemic ID and anaemia.

Methods and results Iron content in cardiomyocytes was reduced in Irp-targeted mice. The animals were not anaemic and did not show a phenotype under baseline conditions. Irp-targeted mice, however, were unable to increase left ventricular (LV) systolic function in response to an acute dobutamine challenge. After myocardial infarction, Irp-targeted mice developed more severe LV dysfunction with increased HF mortality. Mechanistically, the activity of the iron–sulphur cluster-containing complex I of the mitochondrial electron transport chain was reduced in left ventricles from Irp-targeted mice. As demonstrated by extracellular flux analysis *in vitro*, mitochondrial respiration was preserved at baseline but failed to increase in response to dobutamine in Irp-targeted cardiomyocytes. As shown by ^{31}P-magnetic resonance spectroscopy *in vivo*, LV phosphocreatine/ATP ratio declined during dobutamine stress in Irp-targeted mice but remained stable in control mice. Intravenous injection of ferric carboxymaltose replenished cardiac iron stores, restored mitochondrial respiratory capacity and inotropic reserve, and attenuated adverse remodelling after myocardial infarction in Irp-targeted mice but not in control mice. As shown by electrophoretic mobility shift assays, IRP activity was significantly reduced in LV tissue samples from patients with advanced HF and reduced LV tissue iron content.

Conclusions ID in cardiomyocytes impairs mitochondrial respiration and adaptation to acute and chronic increases in workload. Iron supplementation restores cardiac energy reserve and function in iron-deficient hearts.

Keywords Iron deficiency • Heart failure • Energy metabolism • Extracellular flux analysis • ^{31}P-Magnetic resonance spectroscopy

*Corresponding author: Tel: +49 (0)511 532-2229, Fax: +49 (0)511 532-3357, Email: kempf.tibor@mh-hannover.de

Open Research Knowledge Graph

Our Objective

- **Contributions Scholarly Knowledge. Structured.**
  - Focus on structuring only ***contributions*** from **natural language processing (NLP) articles**

Our Objective

- **Contributions Scholarly Knowledge. Structured.**
  - Focus on structuring only ***contributions*** from **natural language processing (NLP) articles**

- **Devise an annotation methodology**: NLPContributions

Our Goals

Two-fold:

Our Goals

Two-fold:

1. perform a **pilot annotation exercise to find a systematic set of patterns of subject-predicate-object statements** for the semantic structuring of scholarly contributions that are more or less generically applicable for NLP articles;

Our Goals

Two-fold:

1. perform a **pilot annotation exercise to find a systematic set of patterns of subject-predicate-object statements** for the semantic structuring of scholarly contributions that are more or less generically applicable for NLP articles;

2. ingest the resulting pilot annotated data into the Open Research Knowledge Graph (ORKG) infrastructure as a showcase to **automatically process the digitalized scholarly contribution knowledge elements**.

Our Goals

Two-fold:

1. perform a **pilot annotation exercise to find a systematic set of patterns of subject-predicate-object statements** for the semantic structuring of scholarly contributions that are more or less generically applicable for NLP articles;

2. ingest the resulting pilot annotated data into the Open Research Knowledge Graph (ORKG) infrastructure as a showcase to **automatically process the digitalized scholarly contribution knowledge elements**.
    - The ORKG[1] is a next-generation digital library infrastructure for machine-actionable knowledge content in scholarly articles.

Reference:

1. Jaradeh, Mohamad Yaser, et al. "Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge." *Proceedings of the 10th International Conference on Knowledge Capture*. 2019.

Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

Plan for the Talk

- **NLPContributions Model**
- **The NLPContributions Annotation Guidelines**
- **Pilot Annotated Dataset Characteristics**
- **NLPContributions in the Open Research Knowledge Graph**

NLPContributions Model: Characteristics

- Designed for building a knowledge graph

NLPContributions Model: Characteristics

- Designed for building a knowledge graph
- Not ontologized
    - assumes a bottom-up data-driven design toward ontology discovery

NLPContributions Model: Characteristics

- Designed for building a knowledge graph
- Not ontologized
    - assumes a bottom-up data-driven design toward ontology discovery
- Has a core skeleton model for top-level knowledge systematization.

NLPContributions Model: Characteristics

- Designed for building a knowledge graph
- Not ontologized
    - assumes a bottom-up data-driven design toward ontology discovery
- Has a core skeleton model for top-level knowledge systematization.
    - a root node called Contribution,

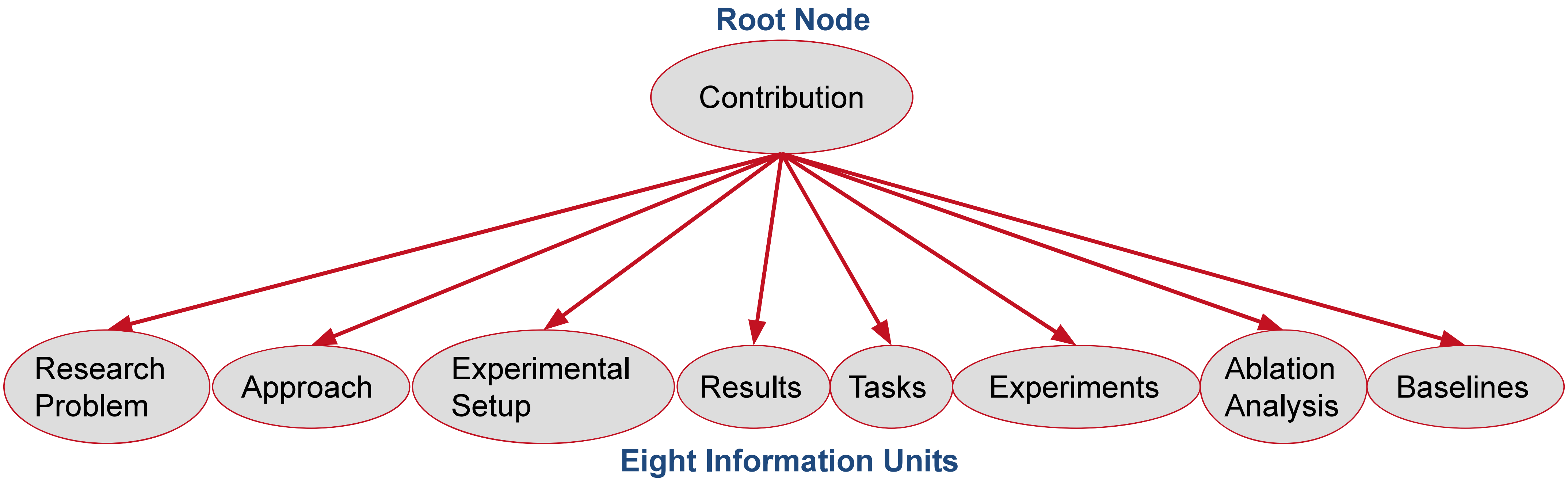
NLPContributions Model: Characteristics

- Designed for building a knowledge graph

- Not ontologized
  - assumes a bottom-up data-driven design toward ontology discovery

- Has a core skeleton model for top-level knowledge systematization.
  - a root node called Contribution,
  - eight first level nodes representing core information units under which the scholarly contributions data is organized
    - inspired from sectional information organization in scholarly articles

NLPContributions Model: Core Skeleton



NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  1. ResearchProblem
  2. Approach
  3. ExperimentalSetup
  4. Results
  5. Tasks
  6. Experiments
  7. AblationAnalysis
  8. Baselines

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  1. ResearchProblem

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  1. ResearchProblem

  - research challenge addressed by a contribution

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  1. ResearchProblem

  - research challenge addressed by a contribution
  - connected to root by predicate *hasResearchProblem*

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  1. ResearchProblem

  - research challenge addressed by a contribution
  - connected to root by predicate *hasResearchProblem*
  - E.g., from paper about BioBERT word embeddings, their research problem is ‘domain-customization of BERT’

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  1. ResearchProblem

  - research challenge addressed by a contribution
  - connected to root by predicate *hasResearchProblem*
  - E.g., from paper about BioBERT word embeddings, their research problem is ‘domain-customization of BERT’
  - typically found in an article’s Title, Abstract and first few paragraphs of the Introduction

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  1. ResearchProblem

  - research challenge addressed by a contribution
  - connected to root by predicate *hasResearchProblem*
  - E.g., from paper about BioBERT word embeddings, their research problem is ‘domain-customization of BERT’
  - typically found in an article’s Title, Abstract and first few paragraphs of the Introduction
  - involves annotating one or more sentences and precisely the research problem phrase boundaries in the sentences

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  2. Approach

  - solution proposed for the research problem

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  2. Approach

  - solution proposed for the research problem
  - connected to root by predicate *has*

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

2. Approach

  - solution proposed for the research problem
  - connected to root by predicate *has*
  - alternatively called Model or Method or Architecture or System or Application

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

2. Approach

- solution proposed for the research problem
- connected to root by predicate *has*
- alternatively called Model or Method or Architecture or System or Application
- typically found in the article's Introduction section in the context of cue phrases such as "we take the approach," "we propose the model," "our system architecture," or "the method proposed in this paper."
  - exception: the first few lines within the main system description content in the article

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  3. ExperimentalSetup

  - details about the platform including both hardware (e.g., GPU) and software (e.g., Tensorflow library) for implementing the machine learning solution; and of variables, that determine the network structure (e.g., number of hidden units) and how the network is trained (e.g., learning rate), for tuning the software to the task objective

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  3. ExperimentalSetup

  - details about the platform including both hardware (e.g., GPU) and software (e.g., Tensorflow library) for implementing the machine learning solution; and of variables, that determine the network structure (e.g., number of hidden units) and how the network is trained (e.g., learning rate), for tuning the software to the task objective
  - connected to root by predicate *has*

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  3. ExperimentalSetup

  - details about the platform including both hardware (e.g., GPU) and software (e.g., Tensorflow library) for implementing the machine learning solution; and of variables, that determine the network structure (e.g., number of hidden units) and how the network is trained (e.g., learning rate), for tuning the software to the task objective
  - connected to root by predicate *has*
  - found in the sections called Experiment, Experimental Setup, Implementation, Hyperparameters, or Training

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  4. Results

  - main findings or outcomes reported in the article for the research problem

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  4. Results

  - main findings or outcomes reported in the article for the research problem
  - connected to root by predicate *has*

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

4. Results

- main findings or outcomes reported in the article for the research problem
- connected to root by predicate *has*
- found in an article's Results, Experiments, or Tasks sections
    - while the results are often highlighted in the Introduction, unlike the Approach unit, in this case, we annotate the dedicated, detailed section on Results because results constitute a primary aspect of the contribution.

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  5. Tasks

  - the Approach, particularly in multi-task settings, are tested on more than one task, in which case, all the experimental tasks are listed

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  5. Tasks

  - the Approach, particularly in multi-task settings, are tested on more than one task, in which case, all the experimental tasks are listed
  - connected to root by predicate *has*

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

5. Tasks

- the Approach, particularly in multi-task settings, are tested on more than one task, in which case, all the experimental tasks are listed
- connected to root by predicate *has*
- is an encapsulating information unit
  - can include one or more of the ExperimentalSetup, Hyperparameters, and Results as sub information units

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

1. ResearchProblem
2. Approach
3. ExperimentalSetup
4. Results
5. Tasks
6. **Experiments**
7. **AblationAnalysis**
8. **Baselines**

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  6. Experiments

  - is an encapsulating information unit
    - can be a combination of ExperimentalSetup and Results; or lists of Tasks and their Results; or Approach, ExperimentalSetup and Results combined

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

6. Experiments

- is an encapsulating information unit
  - can be a combination of ExperimentalSetup and Results; or lists of Tasks and their Results; or Approach, ExperimentalSetup and Results combined
- particularly relevant in the content of multitask systems such as BERT
  - modeling ExperimentalSetup with Results or Tasks with Results is necessary in such systems since the experimental setup often changes per task producing a different set of results

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

7. AblationAnalysis

  - describes the performance of components in systems

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  7. AblationAnalysis

  - describes the performance of components in systems
  - a form of the results which are relevant to a Contribution

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  7. AblationAnalysis

  - describes the performance of components in systems
  - a form of the results which are relevant to a Contribution
  - typically found in sections with Ablation in the title, otherwise also in the running text

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

  8. Baselines

  - a list of systems that a proposed approach is compared against

NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

8. Baselines

- a list of systems that a proposed approach is compared against
- a form of the results which are relevant to a Contribution

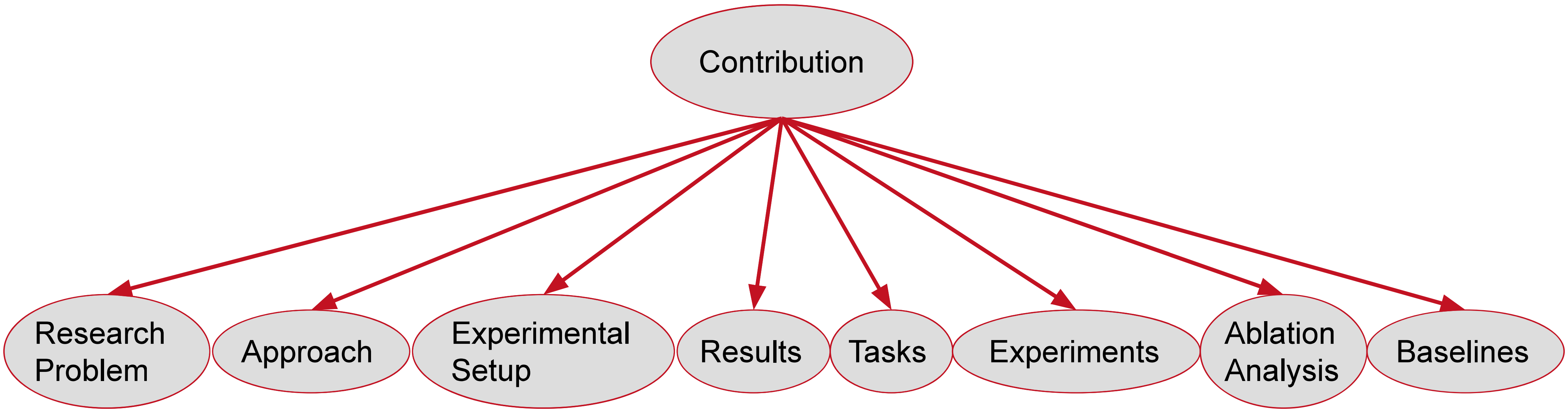
NLPContributions Model: 8 Information Units

- Inspired from sectional information organization in scholarly articles

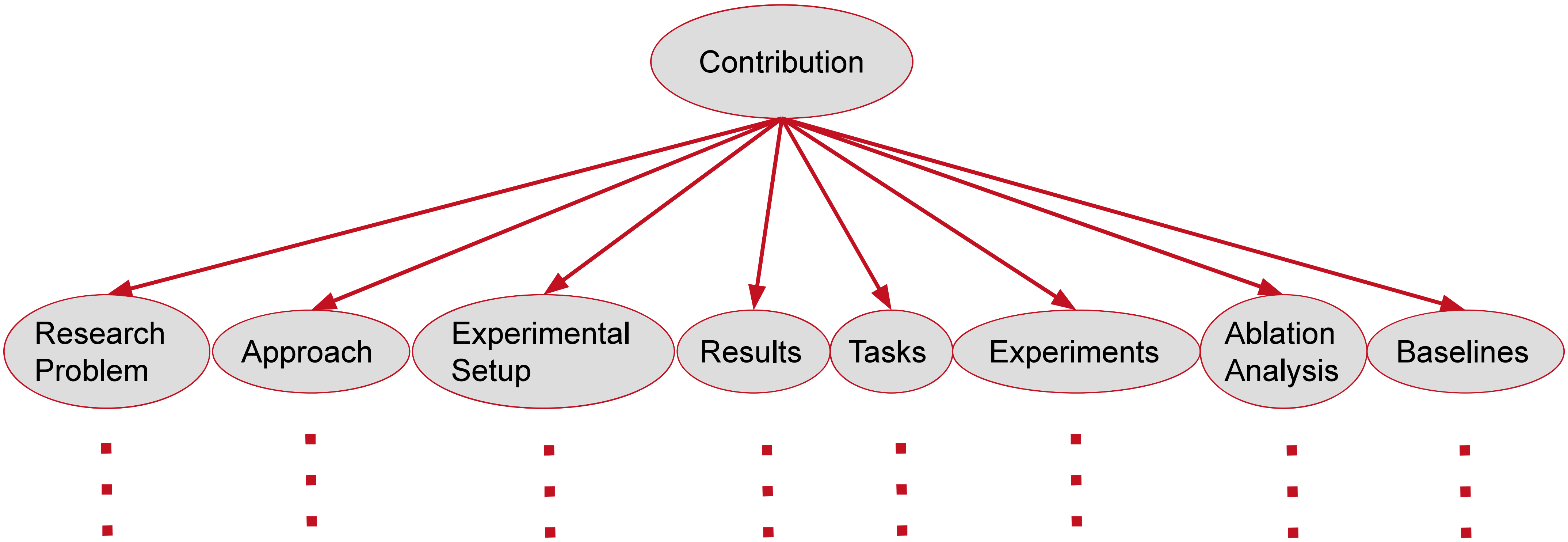
8. Baselines

- a list of systems that a proposed approach is compared against
- a form of the results which are relevant to a Contribution
- typically found in sections with Baseline in the title, otherwise also in the running text

NLPContributions Model: 8 Information Units



NLPContributions Model: <u>Data Elements</u>



How to: Knowledge Graph building from Unstructured Text

- Given a paragraph(s) of unstructured text
  - identify the elements to model:
    - depends on:
      1. if the knowledge graph has an overarching knowledge theme
      2. or, if the knowledge nodes are to be of a certain type (e.g., scientific entities)
    - 1 subsumes 2
  - For 1 (our contributions-themed model):
    - identify the sentences that reflect the theme
    - identify the knowledge entities and predicates from the sentence of interest to the knowledge theme (e.g., scientific entities)
    - create subject-predicate-object triples toward RDFized KGs
    - ...

NLPContributions Model: Data Elements

NLPContributions Model: Data Elements

- **Contribution Sentences**
    - select candidate contribution sentences under each of the aforementioned 3 or more applicable information units (viz., ResearchProblem, Approach, Results, AblationAnalysis, etc.).

NLPContributions Model: Data Elements

- **Contribution Sentences**
  - select candidate contribution sentences under each of the aforementioned 3 or more applicable information units (viz., ResearchProblem, Approach, Results, AblationAnalysis, etc.).
- **Scientific Term and Predicate Phrases as Knowledge Entities (Graph Nodes)**
  - select phrases with an implicit understanding of whether they take the subject, predicate, or object roles in a per-triple context

NLPContributions Model: Data Elements

- **Contribution Sentences**
  - select candidate contribution sentences under each of the aforementioned 3 or more applicable information units (viz., ResearchProblem, Approach, Results, AblationAnalysis, etc.).
- **Scientific Term and Predicate Phrases as Knowledge Entities (Graph Nodes)**
  - select phrases with an implicit understanding of whether they take the subject, predicate, or object roles in a per-triple context
- **Create Triples in Contribution Sequences**
  - relating phrases in subject, predicate, and object roles within triples
  - creating contribution sequences by using an object in one triple as the subject in another triple

NLPContributions Model: Data Elements

Next: Example modeling data elements under an information unit

NLPContributions Model: Approach Data Elements


```
{
  "has" : {
    "Approach" : {
      "converting questions" : {
        "to (un-interpretable) vectorial representations" : {
          "which require" : "no pre-defined grammars or lexicons",
          "can query" : {
            "any KB" : {
              "independent of" : "schema"
            }
          }
        },
        "from sentence" : "In this paper, we instead take the
          approach of converting questions to (un-interpretable)
          vectorial representations which require no pre-defined
          grammars or lexicons and can query any KB independent of
          its schema."
      }
    }
  }
}
```


Reference: Bordes, Antoine, Jason Weston, and Nicolas Usunier. "Open question answering with weakly supervised embedding models." *Joint European conference on machine learning and knowledge discovery in databases*. Springer, Berlin, Heidelberg, 2014.

NLPContributions Model: ExperimentalSetup Data Elements


```
{
  "has" : {
    "Experimental setup" : {
      "used" : [
        {
        "BERTBase model" : {
          "pre-trained for" : "1M steps",
          "pre-trained on" : ["English Wikipedia",
            "BooksCorpus"]
        },
        "from sentence" : "We used the BERTBASE model
          pre-trained on English Wikipedia and
          BooksCorpus for 1M steps."
        },
        {
        "NVIDIA V100 (32GB) GPUs" : {
          "used" : {
            "eight" : {
              "for" : "pre-training"
            }
          },
          "from sentence" : "We used eight NVIDIA V100
            (32GB) GPUs for the pre-training."
          }
        }
      ]
    }
  }
}
```


Reference: Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

NLPContributions Model: Result Data Elements


```
{
  "CoNLL test set" : {
    "for" : {
      "NER" : {
        "F1-score" : "91.57%"
      }
    },
    "from sentence" : "For NER (Table 7), S-LSTM
      gives an F1-score of 91.57% on the CoNLL
      test set, which is significantly better
      compared with BiLSTMs."
  }
}
```


Reference: Zhang, Yue, Qi Liu, and Linfeng Song. "Sentence-State LSTM for Text Representation." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.

Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

NLPContributions Annotation Guidelines

Three of Twelve Guidelines:

NLPContributions Annotation Guidelines

Three of Twelve Guidelines:

1. *How are information unit names selected?* or conversely, *Which of the eight information units does the sentence belong to?*
    - applied name is the one selected based on the closest section title or cue phrase

NLPContributions Annotation Guidelines

Three of Twelve Guidelines:

1. *How are information unit names selected?* or conversely, *Which of the eight information units does the sentence belong to?*
   - applied name is the one selected based on the closest section title or cue phrase
2. *Inferring Predicates*
   - from running text or from the closed class set {"has", "on", "by", "for", "has value", "has description", "based on", "called"}

NLPContributions Annotation Guidelines

Three of Twelve Guidelines:

1. *How are information unit names selected?* or conversely, *Which of the eight information units does the sentence belong to?*
   - applied name is the one selected based on the closest section title or cue phrase
2. *Inferring Predicates*
   - from running text or from the closed class set {“has”, “on”, “by”, “for”, “has value”, “has description”, “based on”, “called”}
3. *How are lists modeled within contribution sequences?*
   - list items are treated just as sentences

Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

Pilot Annotated Dataset

- **Dataset**
  - A collection of scholarly articles downloaded from https://paperswithcode.com/
    - represents papers in AI at large

Pilot Annotated Dataset

- **Dataset**
  - A collection of scholarly articles downloaded from https://paperswithcode.com/
    - represents papers in AI at large
  - Randomly selected 50 NLP papers
    - Aim: create a representative dataset
    - select a distribution of 10 papers across five different NLP research tasks:
      - machine translation, named entity recognition, question answering, relation classification, and text classification.

Pilot Annotated Dataset

- **Dataset**
  - A collection of scholarly articles downloaded from https://paperswithcode.com/
    - represents papers in AI at large
  - Randomly selected 50 NLP papers
    - Aim: create a representative dataset
    - select a distribution of 10 papers across five different NLP research tasks:
      - machine translation, named entity recognition, question answering, relation classification, and text classification.
- **Annotation Tools**
  - https://jsoneditoronline.org/ - For JSON syntax checks
  - https://www.orkg.org/ - As a litmus test for contributions-themed KG and as the Digital Library infrastructure to populate with the annotated KGs

Pilot Annotated Dataset Characteristics

- Total of 2631 triples (avg. of 52 triples per article)
- Data elements: 1033 unique subjects, 843 unique predicates, and 2182 unique objects

Pilot Annotated Dataset Characteristics

- Total of 2631 triples (avg. of 52 triples per article)
- Data elements: 1033 unique subjects, 843 unique predicates, and 2182 unique objects

	MT	NER	QA	RC	TC
Subject	259	209	203	228	221
Predicate	243	220	187	201	252
Object	471	434	515	455	459
Total	502	473	497	544	504

MT: machine translation; NER: named entity recognition; QA: question answering; RC: relation classification; TC: text classification

Pilot Annotated Dataset Characteristics

- Total of 2631 triples (avg. of 52 triples per article)
- Data elements: 1033 unique subjects, 843 unique predicates, and 2182 unique objects

	MT	NER	QA	RC	TC
Subject	259	209	203	228	221
Predicate	243	220	187	201	252
Object	**471**	**434**	**515**	**455**	**459**
Total	502	473	497	544	504

MT: machine translation; NER: named entity recognition; QA: question answering; RC: relation classification; TC: text classification

Pilot Annotated Dataset Characteristics

- Total of 2631 triples (avg. of 52 triples per article)
- Data elements: 1033 unique subjects, 843 unique predicates, and 2182 unique objects

	MT	NER	QA	RC	TC
Subject	259	209	203	228	221
Predicate	243	220	187	201	252
Object	471	434	515	455	459
Total	502	473	497	**544**	504

MT: machine translation; NER: named entity recognition; QA: question answering; RC: relation classification; TC: text classification

Pilot Annotated Dataset Characteristics

- Total of 2631 triples (avg. of 52 triples per article)
- Data elements: 1033 unique subjects, 843 unique predicates, and 2182 unique objects

	MT	NER	QA	RC	TC
Subject	259	209	203	228	221
Predicate	243	220	187	201	252
Object	471	434	515	455	459
Total	502	**473**	497	544	504

MT: machine translation; NER: named entity recognition; QA: question answering; RC: relation classification; TC: text classification

Pilot Annotated Dataset Characteristics

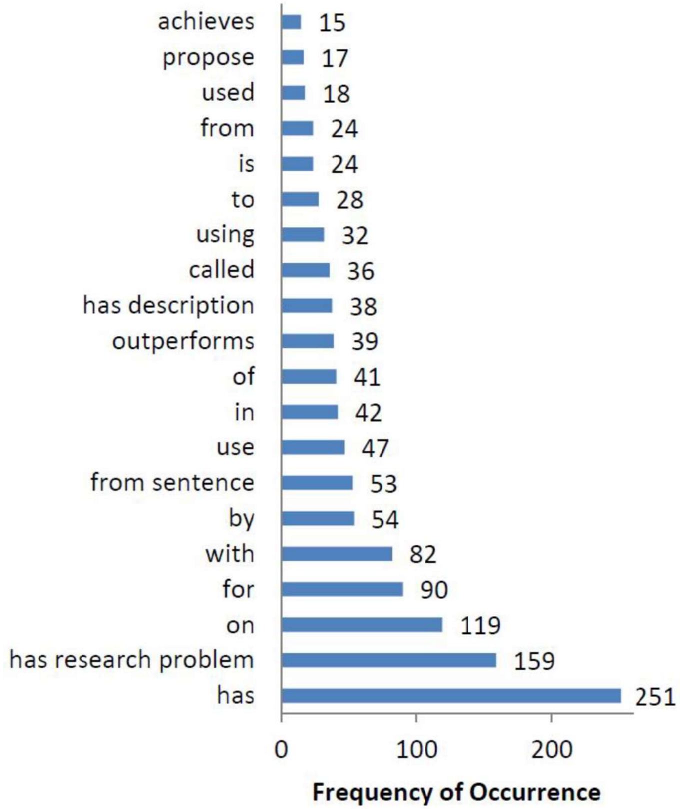
- Total of 2631 triples (avg. of 52 triples per article)
- Data elements: 1033 unique subjects, 843 unique predicates, and 2182 unique objects

	MT	NER	QA	RC	TC
Subject	259	209	203	228	221
Predicate	243	220	187	201	252
Object	471	434	515	455	459
Total	502	473	497	544	504

MT: machine translation; NER: named entity recognition; QA: question answering; RC: relation classification; TC: text classification

Pilot Annotated Dataset Characteristics

- Total of 2631 triples (avg. of 52 triples per article)
- Data elements: 1033 unique subjects, 843 unique predicates, and 2182 unique objects



Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

https://arxiv.org/abs/1809.10185 80%

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org > cs > arXiv:1809.10185

Search... All fields Search

Help | Advanced Search

Computer Science > Computation and Language

[Submitted on 26 Sep 2018]

Graph Convolution over Pruned Dependency Trees Improves Relation Extraction

Yuhao Zhang, Peng Qi, Christopher D. Manning

Dependency trees help relation extraction models capture long-range relations between words. However, existing dependency-based models either neglect crucial information (e.g., negation) by pruning the dependency trees too aggressively, or are computationally inefficient because it is difficult to parallelize over different tree structures. We propose an extension of graph convolutional networks that is tailored for relation extraction, which pools information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant content, we further apply a novel pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities among which a relation might hold. The resulting model achieves state-of-the-art performance on the large-scale TACRED dataset, outperforming existing sequence and dependency-based neural models. We also show through detailed analysis that this model has complementary strengths to sequence models, and combining them further improves the state of the art.

Comments: EMNLP 2018. Code available at: this https URL
Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:1809.10185 **[cs.CL]**
(or arXiv:1809.10185v1 **[cs.CL]** for this version)

Bibliographic data

[Enable Bibex (What is Bibex?)]

Submission history

From: Yuhao Zhang [view email]
[v1] Wed, 26 Sep 2018 18:49:07 UTC (411 KB)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Download:

- PDF
- Other formats

(license)

Current browse context:
cs.CL
< prev | next >
new | recent | 1809
Change to browse by:
cs

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

DBLP - CS Bibliography

listing | bibtex
Yuhao Zhang
Peng Qi
Christopher D. Manning

Export citation

Bookmark

Browse v0.3.2.5 released 2020-07-27 Feedback?

TIB

Abstract

Dependency trees help relation extraction models capture long-range relations between words. However, existing dependency-based models either neglect crucial information (e.g., negation) by pruning the dependency trees too aggressively, or are computationally inefficient because it is difficult to parallelize over different tree structures. We propose an extension of graph convolutional networks that is tailored for relation extraction, which pools information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant content, we further apply a novel pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities among which a relation might hold. The resulting model achieves state-of-the-art performance on the large-scale TACRED dataset, outperforming existing sequence and dependency-based neural models. We also show through detailed analysis that this model has complementary strengths to sequence models, and combining them further improves the state of the art.

Graph Convolution over Pruned Dependency Trees Improves Relation Extraction

2018 | Information Science | Yuhao Zhang | Peng Qi | Christopher D Manning

Published in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*

Contribution 1

Research problems

Relation extraction

Add to comparison

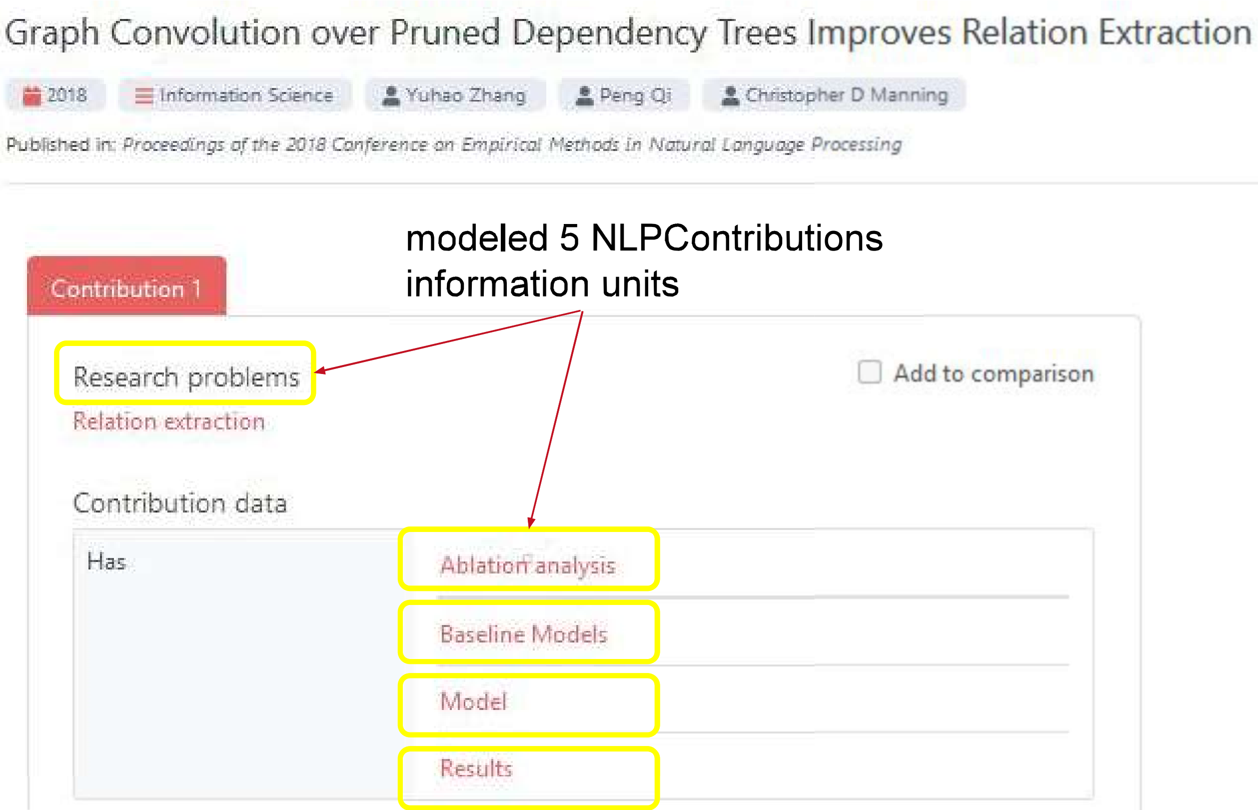
Contribution data

Has	Ablation analysis
	Baseline Models
	Model
	Results

Accessible at https://www.orkg.org/orkg/paper/R44287

Abstract

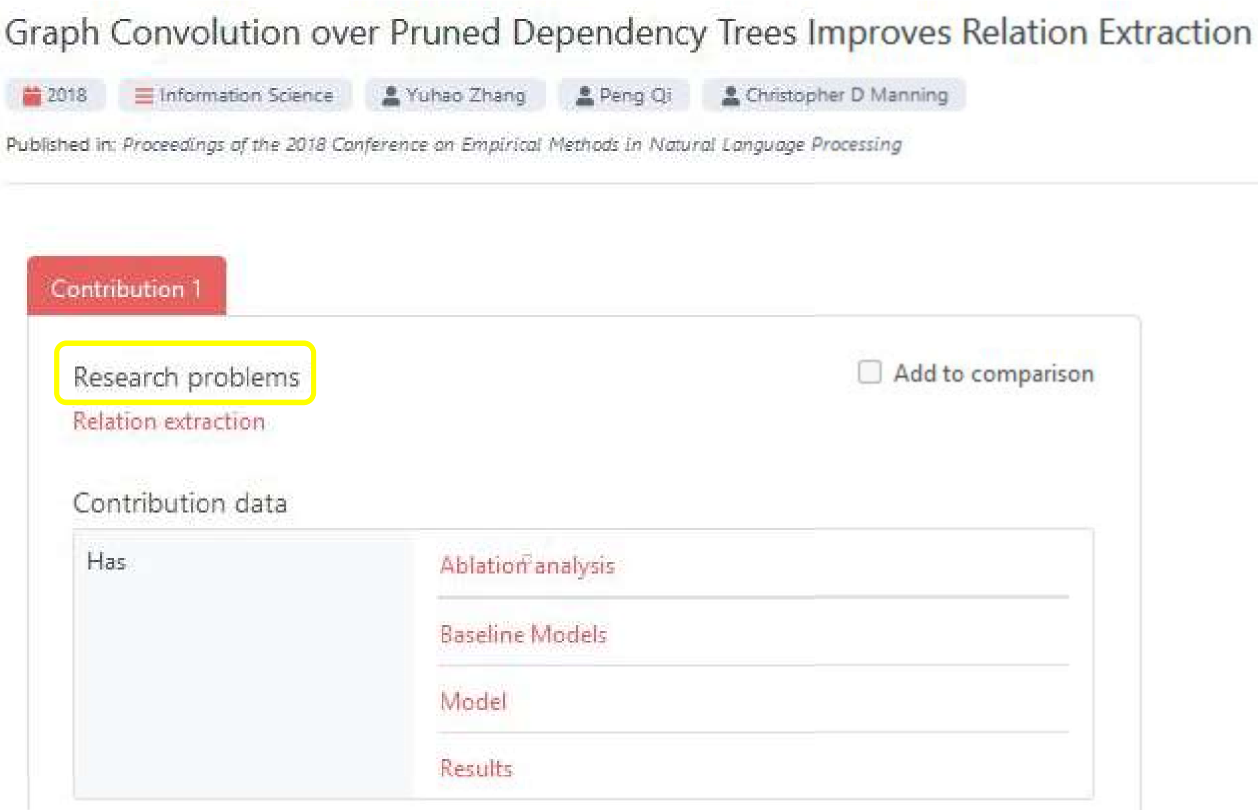
Dependency trees help relation extraction models capture long-range relations between words. However, existing dependency-based models either neglect crucial information (e.g., negation) by pruning the dependency trees too aggressively, or are computationally inefficient because it is difficult to parallelize over different tree structures. We propose an extension of graph convolutional networks that is tailored for relation extraction, which pools information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant content, we further apply a novel pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities among which a relation might hold. The resulting model achieves state-of-the-art performance on the large-scale TACRED dataset, outperforming existing sequence and dependency-based neural models. We also show through detailed analysis that this model has complementary strengths to sequence models, and combining them further improves the state of the art.



Accessible at https://www.orkg.org/orkg/paper/R44287

Abstract

Dependency trees help relation extraction models capture long-range relations between words. However, existing dependency-based models either neglect crucial information (e.g., negation) by pruning the dependency trees too aggressively, or are computationally inefficient because it is difficult to parallelize over different tree structures. We propose an extension of graph convolutional networks that is tailored for relation extraction, which pools information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant content, we further apply a novel pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities among which a relation might hold. The resulting model achieves state-of-the-art performance on the large-scale TACRED dataset, outperforming existing sequence and dependency-based neural models. We also show through detailed analysis that this model has complementary strengths to sequence models, and combining them further improves the state of the art.



Accessible at https://www.orkg.org/orkg/paper/R44287

TIB

Abstract

Dependency trees help relation extraction models capture long-range relations between words. However, existing dependency-based models either neglect crucial information (e.g., negation) by pruning the dependency trees too aggressively, or are computationally inefficient because it is difficult to parallelize over different tree structures. We propose an extension of graph convolutional networks that is tailored for relation extraction, which pools information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant content, we further apply a novel pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities among which a relation might hold. The resulting model achieves state-of-the-art performance on the large-scale TACRED dataset, outperforming existing sequence and dependency-based neural models. We also show through detailed analysis that this model has complementary strengths to sequence models, and combining them further improves the state of the art.

Graph Convolution over Pruned Dependency Trees Improves Relation Extraction

2018 | Information Science | Yuhao Zhang | Peng Qi | Christopher D Manning

Published in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*

Contribution 1

Research problems

Relation extraction

Add to comparison

Contribution data

Has	Ablation analysis
	Baseline Models
	Model
	Results

Accessible at https://www.orkg.org/orkg/paper/R44287

Abstract

Dependency trees help relation extraction models capture long-range relations between words. However, existing dependency-based models either neglect crucial information (e.g., negation) by pruning the dependency trees too aggressively, or are computationally inefficient because it is difficult to parallelize over different tree structures. We propose an extension of graph convolutional networks that is tailored for relation extraction, which pools information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant content, we further apply a novel pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities among which a relation might hold. The resulting model achieves state-of-the-art performance on the large-scale TACRED dataset, outperforming existing sequence and dependency-based neural models. We also show through detailed analysis that this model has complementary strengths to sequence models, and combining them further improves the state of the art.

Graph Convolution over Pruned Dependency Trees Improves Relation Extraction

2018 | Information Science | Yuhao Zhang | Peng Qi | Christopher D Manning

Published in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*

Contribution 1

Research problems

Relation extraction

Add to comparison

Contribution data

Has	Ablation analysis
	Baseline Models
	Model
	Results

Accessible at https://www.orkg.org/orkg/paper/R44287

Plan for the Talk

- NLPContributions Model
- The NLPContributions Annotation Guidelines
- Pilot Annotated Dataset Characteristics
- NLPContributions in the Open Research Knowledge Graph

Conclusion: Takeaways

- Scholarly work can be realized as expressions other than an article
    - We proposed the **NLPContributions annotation model to create contributions-themed knowledge graphs**

Conclusion: Takeaways

- Scholarly work can be realized as expressions other than an article
  - We proposed the **NLPContributions annotation model to create contributions-themed knowledge graphs**
- In a pilot annotation exercise we have annotated 50 articles by the NLPContributions scheme as a practical demonstration of feasibility of the annotation task
  - Available online at https://doi.org/10.25835/0019761

Conclusion: Takeaways

- Scholarly work can be realized as expressions other than an article
  - We proposed the **NLPContributions annotation model to create contributions-themed knowledge graphs**
- In a pilot annotation exercise we have annotated 50 articles by the NLPContributions scheme as a practical demonstration of feasibility of the annotation task
  - Available online at https://doi.org/10.25835/0019761
- The NLPContributions annotation scheme can be leveraged to annotate a larger dataset (of a few hundreds of articles)
  - Train machine-learning-based automated machine readers to annotate tens of thousands of articles for contributions-based KG data which is humanly impossible to do

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

Thank you for your attention!

Questions?

Leibniz
Gemeinschaft