Information Extraction from Texts Using Heterogeneous Information

Makoto Miwa$^{1,2}$
makoto-miwa@toyota-ti.ac.jp

$^1$Computational Intelligence Laboratory, Toyota Technological Institute

$^2$Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST)

Some results in this talk are joint work with the National Centre for Text Mining (NaCTeM), The University of Manchester
Agenda

• Information Extraction
  • Named Entity Recognition (NER)
  • Relation Extraction (RE)
  • Event Extraction (EE)

• Information Extraction Using Heterogeneous Information
  • Document information
  • External information
Agenda

• Information Extraction
  • Named Entity Recognition (NER)
  • Relation Extraction (RE)
  • Event Extraction (EE)

• Information Extraction Using Heterogeneous Information
  • Document information
  • External information
Information Extraction (IE)

The *erbA/myb* IRES virus constructs transformed erythroid cells.

- IE aims at extracting **structured** information from unstructured text
  - We focus on **named entities**, **relations**, and **events**
- Here, I briefly introduce the tasks and our recent models
  - Neural models allow flexible modeling of structures
Nested Named Entity Recognition (Nested NER)

• Named entities are core elements in understanding text

• Traditional entity recognition methods often deal with flat entities, but some recent models consider nested entities
  • E.g., “the premier” is not enough to express the entity
  • Discontinuous entities are also important (e.g., [Zhang et al., 2014])

The premier of the western Canadian province of British Columbia ...
Nested NER Models

• Layered model [Ju et al., 2018]
  • detects entities from inner-most entities to outer entities
  • uses inner entities’ representation for outer entities

• Span model [Sohrab et al., 2018]
  • enumerates all possible regions and classify them into types
  • does not depend on BIO tags

⇔ Neural models allow representing different length spans in the same space
Relation Classification

- A relation is often defined between an entity pair (binary relations).
- Traditional models classify each pair of given entities individually, and they rarely consider their relations, a.k.a., overlapping relations.
- Some inference may be helpful.
  - A Live in B & B Located in C → A Live in C

The premier of the western Canadian province of British Columbia ...
Relation Classification Using Edge-oriented Graphs (EoG) [Christopoulou et al., 2018]

- Aggregates different-length walks on the entity graph for classifying relations
  - Edges have representations unlike graph neural nets
  - Neural models allow aggregation, i.e., representing different-length walks in the same space
Event Extraction

- Events are often represented as **directed acyclic graphs (DAGs)**
- Given edges (binary relations), traditional models build event candidates by enumerating their combinations, and classify them
  - The enumeration is costly and approximation is required
    - Child (or argument) events are often substituted by triggers

```
Pro
Theme
p300 immunoprecipitated

Binding
Theme

Gene expression

Foop3 when both proteins were overexpressed

+Reg

Cause

Theme

Gene expression

+Reg

Cause

Theme2

Theme

in HEK 293T cells
```
Search-based Event Extraction [Kurt et al., 2018]

• searches and fixes events in a bottom-up manner
  • Actions: add, ignore, construct
  • No need to enumerate all the events
  • Child event representations can be used for representing parent events

• maintains **multiple beams** and use **all of them** to find overlapping events

← Neural models allow representing different event structures and entities in the same space
Agenda

• Information extraction tasks
  • Named entity recognition
  • Relation extraction
  • Event extraction

• Information Extraction Using Heterogeneous Information
  • Document information
  • External information
Heterogeneous Background Information for Information Extraction

- **Heterogenous** information (linguistic and non-linguistic) is available to understand text
- Neural models allow representing them in the same/related spaces
- How can we leverage them to improve IE?

**Background:** Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), ...

**Conclusions:** ...

Remdesivir (GS-5734) is an adenosine triphosphate analogue first described in the literature in 2016 as a potential treatment for Ebola. It was superior to placebo in shortening the time to recovery in adults who were hospitalized with Covid-19...

PMID: 32445440

June 11, 2021

BioNLP 2021
Sentence-level Structures

- Syntactic information has been known to be useful for information extraction
  - E.g., shortest path dependency kernels [Bunescu et al., 2005]
- Recent deep models aim to be independent from such information
  - It is not straightforward to incorporate syntactic information into existing deep models

remdesivir was superior to placebo in shortening the time to recovery in adults who were hospitalized with Covid-19...
Syntactically-Informed Word Representation [Tran et al., 2020]

• Inject syntactic information (POS, dependencies) into embeddings using graph convolutional networks
• By just replacing the embeddings layer, existing models benefits from syntactic information

![Graph showing syntactic information integration](image)

<table>
<thead>
<tr>
<th>F-score [%]</th>
<th>NER</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ELMo</td>
<td>SIWR (ELMo)</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>70</td>
<td>65</td>
</tr>
<tr>
<td>70</td>
<td>65</td>
<td>60</td>
</tr>
</tbody>
</table>

NER, RE, F-score [\%]

June 11, 2021  BioNLP 2021
Document-level Information

- Entities are mentioned in a document several times
  - Aggregating information is sometimes helpful to understand entities
- Relations are not always written in a sentence
  ➔ Document-level relation extraction

Background: Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), ...

Conclusions: .. *remdesivir* was superior to placebo in shortening the time to recovery in adults who were hospitalized with Covid-19 ...

Document-level information
Document-level Relation Extraction Using Edge-oriented Graphs (EoGs) [Christopoulou et al., 2019]

- Document-level relations between concepts
- Information is propagated via a document-level graph of mentions, concepts, and sentences
- A concept aggregates the information of their mentions
Iterative Edge Editing for Document-level Relation Extraction

[Makino et al., 2021]

- Iteratively editing edges using relation graphs prebuilt by another system as supports
  - Other edges are used in representing a target edge
  - Encouraging document-level consistency and information propagation
③ Literature Database

• Self-training on raw text corpus
  • Most popular background information in deep learning
    • word embeddings (word2vec [Mikolov et al., 2014]), contextualized embeddings (BERT [Devlin et al., 2019])

• Citation networks are still unexplored for IE, although there are plenty of studies on citation networks

remdesivir was superior to placebo in shortening the time to recovery in adults who were hospitalized with Covid-19...

DeepEventMine: End-to-end Event Extraction [Long et al., 2020]

- Pretrained BERT (SciBERT) as base representation
- Building event structures greedily in a bottom-up manner
  - Corresponding representations are built with simple feed forward neural networks
- SOTA on 7 bio-event corpora
Knowledge Bases

- Knowledge bases contain information on entities and relations
  - Distant supervision is often employed for relation extraction
- Countless efforts and methods for representation learning on knowledge graphs are proposed, but the use of knowledge base information is still limited

remdesivir was superior to placebo in shortening the time to recovery in adults who were hospitalized with Covid-19...
DDI Extraction with Drug Descriptions and Molecule Structures
[Asada et al., 2020]
• Incorporating heterogeneous entity-related information into drug-drug interaction extraction
  • Descriptions: BERT, Molecule structures: Graph neural networks
Relation Extraction with Knowledge Base Priors [Christopoulou et al., 2021]

• Using knowledgebase representations as priors for VAE-based relation extraction model with sentence reconstruction
  • Relation priors are computed from entity pairs by TransE
  • Distant supervision just uses texts matched with knowledge base entries and it does not use entire knowledge bases

![Diagram showing relation extraction process](image.png)
Relation Extraction with Knowledge Base Priors [Christopoulou et al., 2021]

Prior distribution

Posterior distribution

nationality
contains
place_of_birth
place_lived
place_of_death
neighborhood_of
company
admin_divisions
country
capital
Some Thoughts on Information Extraction with Heterogeneous Information

• Neural models dominate recent IE tasks
  • Many deep models work well “without external resources.”
  • Is BERT-like model with more text data enough to perform IE?

• Many questions are still unresolved with external resources
  • What information can we use?
    • This talk misses many, e.g., tables and figures, document attributes, citation networks.
  • How can we utilize multiple, heterogeneous resources?
  • When and how do external resources improve the models?
  • Are there any general way to incorporate external resource information?

• Bio-domain is one of the best domains to organize and investigate external resources
Heterogeneous Phramaceutical Knowledge Graph from DrugBank [Asada et al., 2021]

- Text information in knowledgebase for knowledgebase representations
  - Better entity linking for some links
- We are working on adding more information & using this for IE
Summary

• This talk introduced our recent efforts to information extraction using heterogeneous information

• We investigated and will continue to investigate
  • how to represent IE tasks with neural models
  • how to represent multiple, heterogeneous external information
  • how to combine IE and external information
  • what external information to use

• We are recruiting
  • [https://www.aist.go.jp/aist_e/humanres/ith26e.html](https://www.aist.go.jp/aist_e/humanres/ith26e.html)
References


References


• Thomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. NIPS 2014.
