Multilingual Information Extraction to Learn Terminological Concept Systems

Dagmar Gromann, Centre for Translation Studies, University of Vienna
dagmar.gromann@univie.ac.at

Cardiff NLP Seminar, 18 March 2021
Overview

TERMINOLOGY AND TERMINOLOGICAL CONCEPT SYSTEMS (TCS)

TERM EXTRACTION

RELATION EXTRACTION

SENTENCE-LEVEL TCS LEARNING

DOCUMENT-LEVEL TCS LEARNING
Terminology

- as a field: multidisciplinary field of study that borrows from logic, epistemology, linguistics, philosophy, translation studies, and cognitive science

- as a resource:
  - collection of concepts, their interrelations and designations in a specialized field
  - multilingual designations in a specific subject field structured by concepts, i.e. domain-specific
  - concept? unit of thought, unit of understanding, unit of specialized communication, ...

concept

designation: word name symbol
Terminology

vaccine
Terminological Entry

- Basic model of the Terminological Markup Framework (TMF; ISO 16642: 2017)

- nested internal organization of language information in relation to a concept

```
c_id: 12345
```

```
domain: transportation
```

```
c_id: 12345
de
Fahrrad

en
bicycle

bike
```
Terminological Concept Systems (TCS)

- grouping synonyms and equivalents by concept
- interrelating concepts with hierarchical and non-hierarchical relations

HIERARCHICAL

- generic relation (is_a)
- partitive relation (parts - whole)

NON-HIERARCHICAL

- activity relation (actor - activity, etc.)
- causal relation (cause - effect, etc.)
- ...

Primary functions of a TCS

TCS in principle can be described as structuring means with three major functions (Budin 1996: 18):

1. *epistemic*: epistemological instrument in the sense of structuring knowledge (acquire new knowledge)

2. *informational*: structuring means for practical knowledge transfer (structuring existing knowledge)

3. *communicative*: optimization of specialized communication in the sense of communication organization and consistency (extend existing knowledge)
Related Fields

- Ontology learning: e.g. Petrucci et al. (2018) utilize Neural Machine Translation (NMT)

- Entity extraction and linking:
  - interlinking named entities with semantic, (non-)hierarchical relations
  - mostly on sentence-level, some exceptions on document-level (e.g. DocRED by Yao et al. 2019)
Extracting Terminological Concept Systems from Natural Language Text (Text2TCS)

**OBJECTIVE**

foster terminological consistency to avoid misunderstandings

**METHOD**

ontology learning

machine learning

information extraction

**TCS**

multilingual

hierarchical and

non-hierarchical

relations

**Text2TCS**

Develop a language technology to automatically extract Terminological Concept Systems from multilingual text

**Text2TCS**
Text2TCS Components

GUI

Plain Text

Text File(s)

Text corpora

RDF / TBX

TCS Visualization

ELG API

<<component>> Language recognition

<<component>> Term extraction

<<component>> Synonymy identification

<<component>> Relation extraction

Language

TermCandidates

TermGroups
Initial Idea

DRS

TCS

FRED
Machine Reading for the Semantic Web

http://wit.istc.cnr.it/stlab-tools/fred/
Initial Idea

**DRS**

- \( x_1 \)
- \( \text{named}(x_1, \text{john}, \text{per}) \)
- \( s_2 \ x_3 \)
- \( \text{go}(s_2) \)
- \( \text{agent}(s_2, x_1) \)
- \( \text{school}(x_3) \)
- \( \text{to}(s_2, x_3) \)

**TCS**

**TCS in German, French, etc.**
Pretrained Multilingual Language Models

Bidirectional Encoder Representations from Transformers (BERT)

LM-based TCS Learning

- Pretrained Language Models (LMs) can easily be adapted to a specific task

- Task at hand: Given a sentence identify all domain-specific terms and relations between them

- Challenges:
  - n-gram length that combines to form a term?
  - domain-specific, e.g. vaccine vs. cross-domain-specific, e.g. statistics - both should be extracted
  - Restrictions on sequence length and ability to ensure that two related terms occur in the same input sequence - difficult across sentences but also within long sentences
  - enable this task across many different languages
We meta-analyzed mortality using random-effect models. We meta-analyzed mortality using random-effect models.
We meta-analyzed mortality using random-effect models.

Results on TermEval

TermEval (Rigouts et al. 2020) was a term extraction challenge building on the ACTER dataset with data in:
- English, French, Dutch
- wind energy and corruption (training), dressage (equitation) (validation), heart failure (test)

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>Sequence Classifier</th>
<th>Token Classifier</th>
<th>NMT</th>
<th>Previous SOTA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F1</td>
<td>Prec</td>
</tr>
<tr>
<td>EN</td>
<td>EN</td>
<td>30.9</td>
<td>84.0</td>
<td>45.2</td>
<td>54.9</td>
</tr>
<tr>
<td>FR</td>
<td>EN</td>
<td>31.1</td>
<td>79.5</td>
<td>44.7</td>
<td>56.7</td>
</tr>
<tr>
<td>NL</td>
<td>EN</td>
<td>22.3</td>
<td>91.1</td>
<td>35.9</td>
<td>55.3</td>
</tr>
<tr>
<td>ALL</td>
<td>EN</td>
<td>31.4</td>
<td>85.8</td>
<td>46.0</td>
<td>54.4</td>
</tr>
<tr>
<td>EN</td>
<td>FR</td>
<td>34.6</td>
<td>79.0</td>
<td>48.1</td>
<td>65.4</td>
</tr>
<tr>
<td>FR</td>
<td>FR</td>
<td>32.2</td>
<td>80.2</td>
<td>46.0</td>
<td>68.7</td>
</tr>
<tr>
<td>NL</td>
<td>FR</td>
<td>26.1</td>
<td>84.7</td>
<td>40.0</td>
<td>62.3</td>
</tr>
<tr>
<td>ALL</td>
<td>FR</td>
<td>33.2</td>
<td>78.9</td>
<td>46.7</td>
<td>62.7</td>
</tr>
<tr>
<td>EN</td>
<td>NL</td>
<td>42.8</td>
<td>89.8</td>
<td>58.0</td>
<td>67.9</td>
</tr>
<tr>
<td>FR</td>
<td>NL</td>
<td>41.3</td>
<td>87.6</td>
<td>56.1</td>
<td>69.2</td>
</tr>
<tr>
<td>NL</td>
<td>NL</td>
<td>32.7</td>
<td>94.1</td>
<td>48.5</td>
<td>71.4</td>
</tr>
<tr>
<td>ALL</td>
<td>NL</td>
<td>40.4</td>
<td>91.5</td>
<td>56.0</td>
<td>70.0</td>
</tr>
</tbody>
</table>
Relation Extraction

causalRelation(COVID-19, cough)

XLM-R

cough. COVID-19. The cough was caused by COVID-19

Prespecified relation typology:

- activity relation (actor - activity, etc.)
- causal relation (cause - effect, etc.)
- generic relation (is_a)
- partitive relation (parts - whole)

...
Evaluation in progress

- Adapting existing datasets to our typology:
  - SemEval 2007 Task 4: Classification of Semantic Relations between Nominals
  - SemEval 2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals
  - WCL hypernym dataset

- Generating new datasets for the task:
  - manually generating TSC from multilingual texts (two experts in German + students in other languages)
  - automatically annotating synonyms in sentences based on patterns, e.g. long form + acronym vs. acronyms without synonyms in sentences
  - manual annotation for negative examples, i.e., no relation to be predicted
Context-Free Relation Extraction

Winning system at the Cognitive Aspects of the Lexicon (CogALex) Shared Task 2020: 4 languages, 3 relations + random


Dagmar Gromann IE for Learning Terminological Concept Systems 17.03.2021
Sentence-Level TCS Learning

Sentence Splitting → Term Extraction → Relation Extraction → TCS building and TBX output
Sentence-Level TCS Learning

Sentence Splitting → Term Extraction → Relation Extraction

TCS building and TBX output

one NMT model, two tasks
Globalized Sentence-Level TCS Learning

Term Extraction

Context-Free Relation Extraction

Sentence-Level Relation Extraction

TCS building and TBX output

Sequence Splitting

synonymy

‘motor vehicle’ refers to any power-driven vehicle, which is normally used for carrying persons or goods by road or for drawing.

... Such vehicles include those used for the carriage of persons or goods.
Challenges and next steps

- very few training datasets for specific relation types, e.g. ownership or developmental relation, also for negative examples, i.e., no relation

- creating TCS data manually
  - time- and human resource-intensive
  - near L1 speakers for all languages to be included
  - domain expertise required on top

- Multilingual but not cross-lingual TCS learning
  - alignment across TCS in different languages
  - handling terminological gaps, e.g. *alunizaje* (en: *ram raid*, de: ???), *Schadenfreude* (..:???)

---

Dagmar Gromann
IE for Learning Terminological Concept Systems
17.03.2021
Conclusion

- structured and high-quality terminologies substantially contribute to specialized multilingual communication as well as translation, localization, etc.

- pretrained multilingual language models are highly performant on term extraction and relation extraction

- joint sentence-level extraction of terms, grouping them to synonyms, and learning their interrelations is still challenging

- Where to go from here?
Thank you for your attention!
References


