Extracting Geographic knowledge from unstructured text

Cillian Berragan

09 March, 2021

1 Methods Overview

1.1 Data collection

☒ Wikipedia
☒ Other (travelblog, geograph)

1.2 Data pre-processing

☒ **Coreference resolution** (Document level)

Headingley is a suburb in Leeds. It is the location of Beckett Park. →
Headingley is a suburb in Leeds. Headingley is the location of Beckett Park.

1.3 Relation Classification Dataset

1. **RoBERTa transformer fine-tuned for token classification**
   - Resolved coreferences
   - Re-label Wikipedia subset using Doccano (as coreferences are resolved)?

   {Headingley} is a {suburb} in {Leeds}.

2. **SemEval 2010 Task 8 style Dataset**
   - Entities automatically identified in (1)
   - Relation type manually labelled in doccano

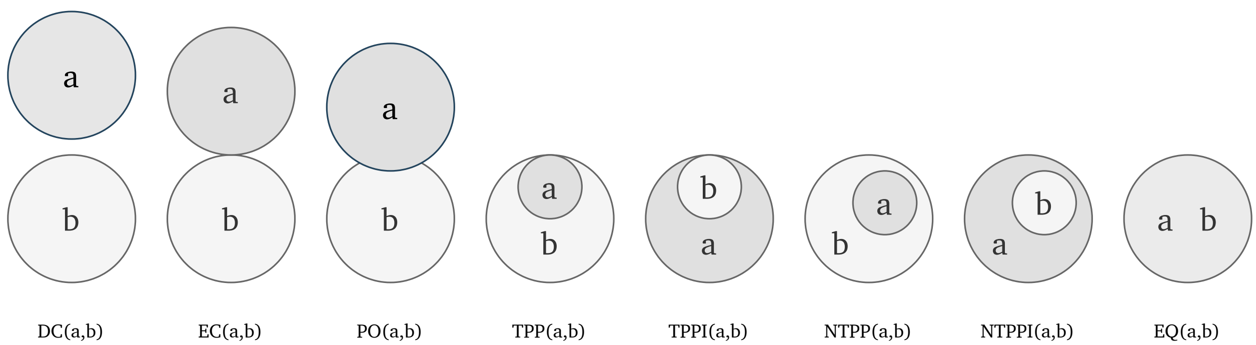
```
Headingley is a suburb of Leeds, West Yorkshire, England, approximately two
miles out of the city centre, to the north west along the A660 road.

Sentence:  [E1]Headingley[/E1] is a suburb in [E2]Leeds[/E2]...
Label:  NTPP(e1,e2)
Sentence:  [E1]Headingley[/E1] is a [E2]suburb[/E2] in Leeds...
Label:  EQ(e1,e2)
Sentence:  [E1]Headingley[/E1] ...  two miles out of the [E2]city centre[/E2]
Label:  DC(e1,e2) (mod:  two miles)
Sentence:  [E1]Headingley[/E1]...  along the [E2]A660[/E2] road.
Label:  EC(e1,e2)
```

NOTE: `eq()` relationships can only exist between a named entity and a nominal. All other relationships must exist between two named entities.

3. **BERT-based relation classification model**
   - R-BERT or Matching the Blanks

1.4 Knowledge graph creation

- ☐ **Known Toponym Resolution**
  - Ordnance Survey open names
  - Mordecai or custom system
  - Wikipedia article coordinates
- ☐ **Knowledge graph**
  - Neo4j
  - TR ⇔ Entity linking
  - Stock (2014): geometric configuration ontology
  - Derive bounding boxes from relations using collection of relations and known place locations



2 Literature

2.1 Coreference

Just problem definition + why needed?

2.2 Geographic Relations

2.2.1 Extraction

(Kordjamshidi *et al.*, 2010; Ludwig *et al.*, 2016; Wallgrün *et al.*, 2014; Qiu *et al.*, 2019)

- Relation Extraction
- Spatial Role Labelling
- Landmark, Trajector, Spatial Indicator
- Various Methods (see Herman (2019)):
  1. Rule-based

2. Weakly supervised
3. Distantly supervised
4. Unsupervised

- Existing work:
    1. GATE Bontcheva *et al.* (2003)
    2. CoreNLP (Stanford)
    3. BERT-based Soares *et al.* (2019), see Soh (2020)

2.2.2 Resolution

(Aflaki & Russell, 2018; Purves *et al.*, 2018; Al-Olimat *et al.*, 2019; Du *et al.*, 2017)

- Region Connection Calculus (RCC8)
- Classifying relations

2.3 Knowledge Graphs

- Geographic knowledge graph as a gazetteer

3 Misc.

3.1 Use of fine-tuned GER model

1. Sequence of tokens $\mathbf{w} = \{w_1, \ldots, w_k\}$ where k is the length of the sequence. Each sequence is present as a sentence in a document $\mathbf{w} \in \mathbf{d}$, and where $\mathbf{d} \in \mathbf{C}$, the collection of all documents. In this case a document is a Wikipedia article summary.
2. Token sequences encoded to integers using a vocabulary of around 50,000 tokens.
3. Passed through transformer layers at various hidden dimensions. Transformer attention capturing context etc.
4. Final RoBERTa layer passed into a linear layer with an output dimension of the number of unique tags t (BILUO tags for PLACE_NAM PLACE_NOM).
5. Outputs: Matrix of logits $\mathbf{L}_{t\times k}$. To find the predicted tags $\mathbf{t} = \{t_1, \ldots, t_k\}$ for a sequence $\mathbf{w}$, the argmax is taken across the k dimension of $\mathbf{L}$ to find the position of the largest *logit* value for each $w_i \in \mathbf{w}$. The values in $\mathbf{t}$ are then mapped to a string representation of the tags.

$$\mathbf{L}_{k,t} = \begin{pmatrix} l_{1,1} & l_{1,2} & \cdots & l_{1,t} \\ l_{2,1} & l_{2,2} & \cdots & l_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ l_{k,1} & l_{k,2} & \cdots & l_{k,t} \end{pmatrix} \xrightarrow{argmax} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} place \\ other \\ \ldots \\ place \end{pmatrix}$$

if $l_{1,1}, l_{2,2}, l_{t,1}$ are the highest *logit* values across the k dimension.

3.2 NOTES

(Ludwig *et al.*, 2016): an issue in extracting spatial semantics from natural language is the lack of annotated data on which machine learning can be employed to learn and extract the spatial relations

Particulars (city county etc.) vs universals instance to universal Liverpool is a city etc.

Note difference between universals and city centre in above sentence

1. Scrap nominals? Instance to instance relationships

NOTES: (Smith, n.d.): Particulars + Universals. Continuants and Occurents.

Formal ontological relation - *part of*.

Three typical top-level relation types: <*universal*, *universal*>: Both relata are universsals. An example of this type of relation is characterization, or the subsumption (*is_a*) relation which obtains between the universtal human and the universal mammal. Such that human is_a mammal.

<**instance**, *universal*>: The first relatum is a particular, the second is a universal. An example of a relation of this type is the instantiation relation, which obtains between this particular person named Peter and thje universal *human*, or between Peter's life and the universal *life*. Another example is the relation of *being allergic to* that exists between Peter and the universal *aspirin*.

<**instance**, **instance**>: Both relata are particulars. Examples include the inherence relation, or the participation relation which obtains between Peter's life and Peter, or also - independently of the ontological sextet - the part-whole relation on the level of instances, which obtains between this particular nose (Peter's nose) and this particular head (Peter's head), and between both of these and peter.

italics for relations between universals and **bold** for relations with at least one particular. Use expressions common in ontology: *is_a* for a subsumption relation and **instance_of** for the instantiation relation. Confined to binary relations.

Knowledge typically concerns itself with universals (e.g. *part_of* relationships) e.g. in Biology.

3.3 In terms of Geography

Particulars, *Universal*

Characterization relation <*universal*, *universal*>: <*suburb*, (*type_of*), *residential area*>

Instantiation relation <**instance**, *universal*>: <**Liverpool**, (**instance_of**), *city*>.

Participation relation <**instance**, **instance**>: <**Liverpool**, (**located_in**), **Merseyside**>

r **instance_of** R

r_1 **located_in** r_2

A *suburb* is a *mixed-use* or *residential area*, existing either as part of a *city* or *urban area* or as a separate *residential community* within commuting distance of a *city*.

<*suburb*, (*type_of*), *mixed-use*>, <*suburb*, (*type_of*), *residential area*>

<*suburb*, (*part_of*), *city*>, <*suburb*, (*part_of*), *urban area*>

<*suburb*, (*near_to*), *city*>

<*suburb*, (*equivalent_to*), *residential community*>

NOTE: part of here is not representing a spatial/geographic relationship. Part of is compositional, a geographic relation would be represented through *located_in*. Suburbs while described as existing as part of a city are not necessarily located within the regional bounds of a city.

Aflaki, N. & Russell, S. (**2018**) *Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions*. p. 6.

Al-Olimat, H.S., Shalin, V.L., Thirunarayan, K. & Sain, J.P. (**2019**) 'Towards Geocoding Spatial Expressions (Vision Paper)', *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '19*. Chicago, IL, USA: ACM Press, 2019, pp. 75–78.

Bontcheva, K., Maynard, D., Tablan, V. & Cunningham, H. (**2003**) *GATE: A Unicode-based Infrastructure Supporting Multilingual Information Extraction*. p. 8.

Du, S., Wang, X., Feng, C.-C. & Zhang, X. (**2017**) Classifying natural-language spatial relation terms with random forest algorithm. *International Journal of Geographical Information Science*. 31 (3). pp. 542–568.

Herman, A. (**2019**) Different ways of doing Relation Extraction from text. *Medium*.

Kordjamshidi, P., Otterlo, M.V. & Moens, M.-F. (**2010**) *Spatial Role Labeling: Task Definition and Annotation Scheme*. p. 8.

Ludwig, O., Liu, X., Kordjamshidi, P. & Moens, M.-F. (**2016**) Deep Embedding for Spatial Role Labeling. *arXiv:1603.08474 [cs]*.

Purves, R.S., Clough, P., Jones, C.B., Hall, M.H. & Murdock, V. (**2018**) *Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text*. now.

Qiu, P., Yu, L., Gao, J. & Lu, F. (**2019**) Detecting geo-relation phrases from web texts for triplet extraction of geographic knowledge: A context-enhanced method. *Big Earth Data*. 3 (3). pp. 297–314.

Smith, B. (**n.d.**) *Ontological relations*.

Soares, L.B., FitzGerald, N., Ling, J. & Kwiatkowski, T. (**2019**) Matching the Blanks: Distributional Similarity for Relation Learning. *arXiv:1906.03158 [cs]*.

Soh, W.T. (**2020**) BERT(S) for Relation Extraction in NLP. *Medium*.

Stock, K. (**2014**) *A Geometric Configuration Ontology to Support Spatial Querying*. p. 8.

Wallgrün, J.O., Klippel, A. & Baldwin, T. (**2014**) 'Building a corpus of spatial relational expressions extracted from web documents', *Proceedings of the 8th Workshop on Geographic Information Retrieval - GIR '14*. Dallas, Texas: ACM Press, 2014, pp. 1–8.