Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

Using LLMs for Enriching Metadata with Links to KOS and Knowledge Graphs: Case Finnish Named Entity Linking

Rafael Leal, Annastiina Ahola ja Eero Hyvönen - 20.10.2024

Overview

- Introduction
- Methods
- Results
- Discussion

Introduction

Rationale

- Metadata enrichment
- Linked Open Data portals
  - parlamenttisampo.fi

Finnish: challenges

- **Declension: ~15 cases**
- **Non-trivial lemmatization**
  - Named entities: basic form is not enough
    - *Suomen korkeimmassa oikeudessa*
- **Lack of open datasets**

Objectives

- **Finnish named entity linker**
  - Increase the digital presence of Finnish
- **Editable results**
- **Database-agnostic**
- **Aware of computational costs**
- **Use open tools**

Methods

Classic 3-step process

- Named entity recognition
- Candidate generation
- Disambiguation and linking

\+ Manual review

Named Entity Recognition

- **Finnish NER tool (2020)**
  - By TurkuNLP Research Group
  - Fine-tuned FinBERT (125M parameters)
  - MIT license

Koripallon lisäksi Jordania kiinnosti moottoriurheilu, ja hän perusti moottoripyöräilytallin, Michael Jordan Motorsportsin. Hän oli mukana mainostamassa monia moottoripyöräkilpailuja.

Candidate generation

- **Lexical matching**
  - Alternative names (redirections)
  - Less robust
  - Easier to import new databases
- **Vectorization**
  - More flexible
  - Harder to maintain
  - Better recall, worse precision

Candidate generation: lemmatization

- **Deep learning**
  - Fine-tuned `Finnish-NLP/t5-small-nl24-finnish`
    - *260M, Apache 2.0*
  - Dataset: ~1M Wikipedia internal links
  - Could be done via a LLM
- **Heuristic lemmatization if needed**

Disambiguation

- **Zero-shot RAG-like**
  - Candidates added to the prompt
  - Wikipedia introductions & Wikidata predicates
- **Llama-3-8B-Instruct**
  - Local LLM with open weights
- **GPT-4 (1-1.8T parameters?)**
  - Closed LLM

Manual review

- Allows end users to review and modify entity links
- UI prototype

Results

Lemmatization

- **Wikipedia internal links**
  - 10K examples
- **Fine-tuned** `Finnish-NLP/t5-small-nl24-finnish` (260M)
  - Finnish only
  - Accuracy: ~96.5%
- **Fine-tuned** `google/mt5-small` (300M)
  - Multi-lingual
  - Accuracy: ~83%

Disambiguation

- **17 automatically retrieved entries**
  - From manually annotated Wikinews dataset
  - Related to Wikipedia disambiguation pages
  - Average of ~5.7 candidates
  - No especially tricky questions
- **Manual evaluation**

- ATV
- Algeria
- Citroën
- Esso
- Gothenburg
- John Roberts
- Kaduna
- NATO
- Odessa (Texas)
- Potamia (Evrytania)
- Progress
- Thames
- Uusi Suomi
- White House

Disambiguation: textual description

- **Llama-3-8B-Instruct**
  - Non-Instruct version could not answer
  - Accuracy ~94% (1 wrong answer)
- **GPT-4**
  - Accuracy 100% (smaller sample)

Disambiguation: YAML

- 5 examples
- Llama-3-8B-Instruct
  - Accuracy 0%
  - Understands the task
- GPT-4
  - Accuracy 100%

(3) Kaduna (joki):
Commons-luokka: Kaduna River
alaluokka kohteelle: liquid water
eri kuin: Kaduna
esiintymä kohteesta: joki
koordinaatit:
- Point(5.801466666 8.741561111):
  pätee osalle: joensuu
- Point(8.7317 9.6911):
  pätee osalle: alkulähde
käyttäjä: maatila
käyttö: merenkulku

Discussion

Future work

- Prompt engineering
- Further model testing (Llama-3.1-70B)

Discussion

- Open-weights LLM?