Using LLMs as Chainsaws – Fostering a Tool-Critical Approach for Information Extraction

language and
translation
technology
team

*Tess Dejaeghere, Pranaydeep Singh, Els Lefever,
Julie Birkholz, Aaron Maladry*

GHENT
UNIVERSITY

Introduction

Who are we?

Pranay

Aaron

Who are we?

8 professors

4 postdocs

18 PhDs

www.lt3.ugent.be

GHENT UNIVERSITY

Ghent Center for Digital Humanities

- Flemish contribution to DARIAH and CLARIN
- unlocking AI tools for research in the Arts and Humanities

CLARIAH-VL
Open Humanities Service Infrastructure

GHENT UNIVERSITY

language and translation technology team

Eu-funded CLS infra

Goal: build a shared resource of high-quality data, tools and knowledge to aid new approaches to studying literature in the digital age

- digital age offers challenges and opportunities for completing research on Europe's multilingual and interconnected literary heritage.
- many resources are currently available in digital libraries, but a lack of standardisation hinders their access and reuse.

=> bridging resource gap

The workshop

GHENT UNIVERSITY

Workshop

- ➔ general introduction to Information Extraction for DH with NLP
  - Information Extraction == What do we want to do?
  - NLP, machine learning == How do we do it?
  - Large Language Models == How do the tools work? ~~ understanding methodology
- ➔ tutorial through notebooks
  - how do we use it? ~~applying methodology
  - three approaches:
    - zero-shot
    - few-shot
    - finetuning / training

Information extraction for DH

GHENT UNIVERSITY

What is information extraction?

broader research question as example:

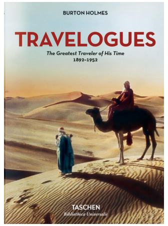
⇒ *What are people talking about in travel literature and what do they say about it?*

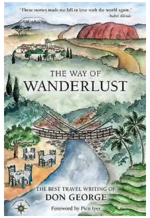
What is information extraction?

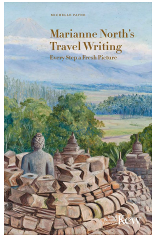
broader research question:

⇒ *What are people talking about in travel literature and what do they say about it?*

Language	18thC	19thC	20thC	Total
English	41	782	668	**1,491**
French	5	145	50	**200**
Dutch	25	92	242	**359**
German	972	218	80	**1,270**
Total	**1,043**	**1,163**	**897**	**3,320**







1. from a corpus of travelogues

GHENT UNIVERSITY

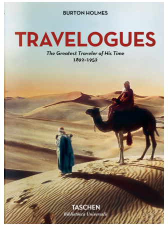
language and translation technology team

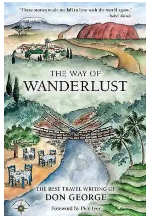
What is information extraction?

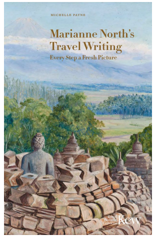
broader research question:

⇒ *What are people talking about in travel literature and what do they say about it?*

Language	18thC	19thC	20thC	Total
English	41	782	668	**1,491**
French	5	145	50	**200**
Dutch	25	92	242	**359**
German	972	218	80	**1,270**
Total	**1,043**	**1,163**	**897**	**3,320**







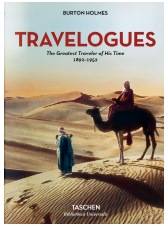
1. from a corpus of travelogues
2. what are people talking about?
    a. named entities

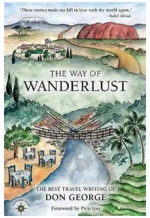
weather	biome	human landform
person	organisation	natural landform
flora	fauna	myth
location		

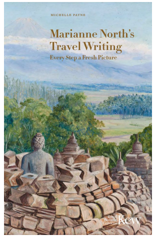
GHENT UNIVERSITY

language and translation technology team

What is information extraction?







broader research question:

⇒ *What are people talking about in travel literature and what do they say about it?*

Language	18thC	19thC	20thC	Total
English	41	782	668	**1,491**
French	5	145	50	**200**
Dutch	25	92	242	**359**
German	972	218	80	**1,270**
Total	**1,043**	**1,163**	**897**	**3,320**

1. from a corpus of travelogues
2. what are people talking about?
   a. named entities
3. what do they say about it?
   a. sentiment analysis

🌥️ weather	⛰️ biome	🏰 human landform
👩 person	🏢 organisation	⛰️ natural landform
🌺 flora	🐱 fauna	🦄 myth
🌍 location		

1 2 3 4 5

Extremely negative 😩 → Extremely positive 😁

GHENT UNIVERSITY

language and translation technology team

What is information extraction

- Task 1: What are people talking about?
  - named entity recognition
- Task 2: What are they saying about it?
  - aspect-based sentiment analysis

⇒ are they positive or negative about these entities?

GHENT UNIVERSITY

language and translation technology team

What is information extraction

- Task 1: What are people talking about?
  - Named entity recognition
- Task 2: What are they saying about it?
  - aspect-based sentiment analysis

⇒ are they positive or negative about these entities?

Widely applicable approach for DH:

⇒ many under-researched books and corpora

GHENT UNIVERSITY

language and translation technology team

Task 1: Named Entity Recognition (NER)

- Which entities are in the text?

> *Rome is a beautiful city, but the Trevi fountain was disappointing.*

Task 1: Named Entity Recognition (NER)

- Which entities are in the text?
  - Rome
  - Trevi fountain

Rome is a beautiful city, but the Trevi fountain was disappointing.

Task 1: Named Entity Recognition (NER)

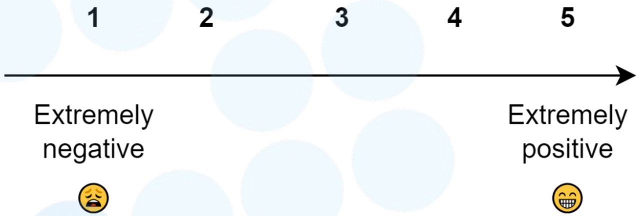
- Which entities are in the text?
  - Rome
  - Trevi fountain
- What types are the named entities?
  - Rome ⇒ location
  - Trevi fountain ⇒ human landform

> *Rome is a beautiful city, but the Trevi fountain was disappointing.*

Task 2: Aspect-Based Sentiment Analysis

- What sentiment is expressed about these entities?
  - Rome: positive
  - Trevi fountain: negative

Rome is a beautiful city, but the Trevi fountain was disappointing.



NLP

GHENT
UNIVERSITY

NLP and Machine Learning

- NLP: Natural Language Processing
  - subfield of linguistics 📝 + computer science 💻
  - large-scale processing of language to answer linguistic questions

GHENT UNIVERSITY
language and translation technology team

NLP and Machine Learning

- NLP: Natural Language Processing
  - subfield of linguistics 📝 + computer science 💻
  - large-scale processing of language to answer linguistic questions
- traditionally: rule based or machine learning

NLP and Machine Learning

- NLP: Natural Language Processing
  - subfield of linguistics 📝 + computer science 💻
  - large-scale processing of language to answer linguistic questions
- traditionally: rule based or machine learning
- currently ⇒ strong focus on machine learning 🤖

Machine Learning 🤖

- machine learning = “giving computers the ability to learn without being explicitly programmed” (Arthur Samuel, 1959).
- not based on linguistic rules, but learning from examples 🧠
  - data-driven

Machine learning

Training data

Pieter B
1 bijdrage

Super interessant en vooral leuk

aug. 2020 • Gezinnen

Super leuke gids die zelf archeologe was geweest
Veel leuke en interessante verhalen met afbeeldingen als extra toelichting. Als toerist is nu de beste tijd om Rome te bezoeken. Nergens wachtrijen en oponthoud.

Machine learning algoritme

Prediction for new data

Rome is a beautiful city, but the Trevi fountain was disappointing.

Rome = location

Trevi Fountain = human landform

Machine learning

Training data

Pieter B
1 bijdrage

Super interessant en vooral leuk
aug. 2020 • Gezinnen
Super leuke gids die zelf archeologe was geweest
Veel leuke en interessante verhalen met afbeeldingen als extra toelichting. Als toerist is nu de beste tijd om Rome te bezoeken. Nergens wachtrijen en oponthoud.

Machine learning algoritme

Prediction for new data

Rome is a beautiful city, but the Trevi fountain was disappointing.

Rome = location 😁
Trevi Fountain = human landform 😩

GHENT UNIVERSITY

language and translation technology team

LLMs

What are LLMs

- machine learning needs to optimize for a task

What are LLMs

- machine learning needs to optimize for a task
- GPT: Generative Pre-trained Transformer

What are LLMs

- machine learning needs to optimize for a task
- GPT: Generative Pre-trained Transformer
- how can a computer generate text?

  👉 one word at a time!

What are LLMs

- machine learning needs to optimize for a task
- GPT: Generative Pre-trained Transformer
- how can a computer generate text?

  👉 one word at a time!

The task:

🤔predict the next word, from left-to-right 🖨️

What are LLMs

- machine learning needs to optimize for a task
- GPT: Generative Pre-trained Transformer
- how can a computer generate text?

  👉 one word at a time!

The task:

🤔predict the next word, from left-to-right 🖨️

→ **causal autoregressive**

What are LLMs

- next-word prediction:
  👌 easy to set up, no supervision needed

There's no place like **<*MASK*>**

This morning I ate a sandwich with peanut butter and **<*MASK*>**

This morning I ate a sandwich with peanut butter and **<*MASK*>**

What are LLMs

- next-word prediction:
  👌 easy to set up, no supervision needed
  🥸 mask the next word (can be done randomly)

There’s no place like **<MASK>**

This morning I ate a sandwich with peanut butter and **<MASK>**

This morning I ate a sandwich with peanut butter and **<MASK>**

What are LLMs

- next-word prediction:
  👌 easy to set up, no supervision needed
  🥸mask the next word (can be done randomly)
  🔮predict which word in the vocabulary is most likely to follow

There's no place like
<MASK>

This morning I ate a
sandwich with peanut butter
and ***<MASK>***

This morning I ate a
sandwich with peanut butter
and ***<MASK>***

What are LLMs

- 🧠✨newest LLMs: two types of training

1. 📖continuous text (pre-training)

There’s no place like <MASK>

This morning I ate a sandwich with peanut butter and <MASK>

This morning I ate a sandwich with peanut butter and <MASK>

GHENT UNIVERSITY

language and translation technology team

What are LLMs

- 🧠✨newest LLMs: two types of training

1. 📖continuous text (pre-training)

There’s no place like <MASK>

This morning I ate a sandwich with peanut butter and <MASK>

This morning I ate a sandwich with peanut butter and <MASK>

2. 👨‍🏫 Instruction tuning for chat models 💬🗨️
   a. also used for fine-tuning

Hey, do you know how to write a script for named entity recognition?

Sure, the first step is to <MASK>

What are LLMs

🤔 how can a computer model language?

🕵️💭Distributional hypothesis: ”You shall know a word by the company it keeps” (Firth, 1957)

What are LLMs

🤔 how can a computer model language?

🕵️💭Distributional hypothesis: ”You shall know a word by the company it keeps” (Firth, 1957)

>> words that occur in similar contexts tend to have related meanings

>> we can infer the meaning of words from context (surrounding words).

GHENT UNIVERSITY language and translation technology team

What are LLMs

🤔 how can a computer model language?

🕵️💭Distributional hypothesis: ”You shall know a word by the company it keeps” (Firth, 1957)

>> words that occur in similar contexts tend to have related meanings

>> we can infer the meaning of words from context (surrounding words).

The national <u>bank</u> of Belgium also gave financial support to the project.

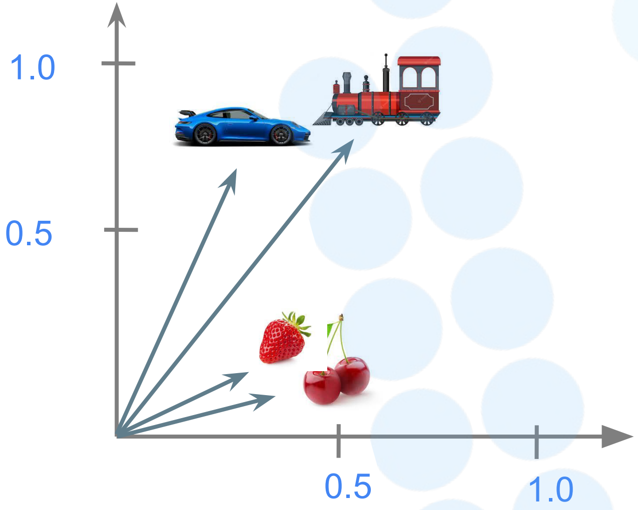
Heavy <u>banks</u> of snow surrounded the train.

Special animals and plants can be found along the <u>banks</u> of the river Meuse.

Modeling word meaning with embeddings

- Computers cannot work with text > we represent words as numeric vectors computers can work with
- Those numbers contain information about the meaning of words, deduced from the contexts in which these words occur (in massive text collections)
- Words that are semantically related (have similar or related meanings) will have similar vectors (and be closer in the vector space)

Embeddings



	0.35	0.15
	0.30	0.20
	0.25	0.70
	0.50	0.75

Embeddings

How does a model know which values to use?

🔢start from random numbers

Embeddings

How does a model know which values to use?

- 🔢start from random numbers
- 🎯trial and error💥
  - 🛠️if the system makes a mistake, change the numbers
  - 🚿rinse and 🔄 repeat
  - 📈 gradually improves

Embeddings

How does a model know which values to use?

- 🔢start from random numbers
- 🎯trial and error💥
  - 🛠️if the system makes a mistake, change the numbers
  - 🚿rinse and 🔄 repeat
  - 📈 gradually improves

🧙almost like magic, this works!

How can this work?

How does a model know which values to use?

Large amount of example data (human-written text)

Training
= trial and error
if the system makes a mistake, change the numbers
repeat

What are LLMs

How big is this model really?

What are LLMs

How big is this model really?

- 📚training data estimate: 25 million books +
- 🦾GPT(3+) model parameters: 1 trillion

🤯models have read more than a human could in a lifetime

LLMs: why is this useful for DH?

🍃🙋🏻‍♂️💬natural human language goes in, natural language comes out

⇒🧾open text

LLMs: why is this useful for DH?

🍃🙋🏻‍♂️💬natural human language goes in, natural language comes out

⇒🧾open text

you ask🧑💬

🗨️🤖LLM answers

GHENT UNIVERSITY

language and translation technology team

LLMs: why is this useful for DH?

🍃🙋🏻‍♂️💬natural human language goes in, natural language comes out
⇒🧾open text

you ask🧑💬
🗨️🤖LLM answers

👨‍🏫give instructions like you would for a human
💡 intuitive, describe ✏️what you want
📈 most probable + human-like response

GHENT UNIVERSITY

language and translation technology team

LLMs: why is this useful for DH?

extensive training

- 🌍many languages, language types
- 📚many existing datasets
- 🚩tasks

idea of foundation models:

- 💱transfer to contexts that are similar to those they are trained for.
- 🏗️ leverage generalization from broad selection of languages and tasks

without needing a lot of data🏷️🗂️

- 🏛️not available for low-resource DH scenarios

How can we use language models?

two types of access:

- 🌐open-source
- 🔐proprietary

Accessing language models: proprietary🔐

- examples
  - Claude
  - GPT 3+
  - Gemini

Accessing language models: proprietary 🔐

- examples
  - Claude
  - GPT 3+
  - Gemini

how to access ⇒ API

Accessing language models: proprietary🔐

- examples
  - Claude
  - GPT 3+
  - Gemini

how to access ⇒ API

✉️you send prompt (your question, instructions) to the company

🏬the company processes⚙️the prompt on their GPU

📨they send you the response

Accessing language models: proprietary🔐

Advantages:

✅availability of huge models

✅no need for GPUs or technical infrastructure

✅works on basic laptop 💻 and internet

Accessing language models: proprietary🔐

Advantages:

✅availability of huge models
✅no need for GPUs or technical infrastructure
✅works on basic laptop 💻 and internet

Disadvantages:

❌cost ! pay for each generated token 💰
❌limited control over the model
- 📉 options for fine-tuning

❌transparency 🙈
- 📚 what pre-training data was used?
- 🚩 which tasks was the model trained on?

Accessing language models: proprietary 🔐

Advantages:

✅availability of huge models
✅no need for GPUs or technical infrastructure
✅works on basic laptop 💻 and internet

Disadvantages:

❌cost ! pay for each generated token 💰
❌limited control over the model
- 📉 options for fine-tuning

❌transparency 🙈
- 📚 what pre-training data was used?
- 🚩 which tasks was the model trained on?

OpenAI

😱SCANDAL😱
😵‍💫optimising models for benchmark data
💥mismatch with real-world performance
🤥"cheating" to present as the best model

GHENT UNIVERSITY

language and translation technology team

Accessing language models: open source 🌐

- examples:
  - Llama
  - Mistral

Meta

MISTRAL
AI_

Accessing language models: open source🌐

- examples:
  - Llama
  - Mistral

- how to access:
  💾download from 🤗Hugging Face
  🛠️run/train them yourself

Meta

MISTRAL
AI_

Accessing language models: open source🌐

Advantages:

✅transparent 😇 about how the model was created

- 📁which data is used
- ⚙️how the data was processed
- 🔧 model architecture

✅can finetune the model yourself

- 🕹️control over the training process
- 🤝 share the model

Accessing language models: open source🌐

Advantages:

✅transparent 😇 about how the model was created

- 📁which data is used
- ⚙️how the data was processed
- 🔧 model architecture

✅can finetune the model yourself

- 🕹️control over the training process
- 🤝 share the model

Disadvantages:

❌require your own GPUs 💎

- 🤏generally limited to smaller models

❌technically more complex 🤓📐

GHENT UNIVERSITY

language and translation technology team

Accessing language models: open source 🌐

Advantages:

✅transparent 😇 about how the model was created

- 📁which data is used
- ⚙️how the data was processed
- 🔧 model architecture

✅can finetune the model yourself

- 🕹️control over the training process
- 🤝 share the model

Disadvantages:

❌require your own GPUs 💎

- 🤏generally limited to smaller models

❌technically more complex 🤓📐

- ☝️BUT we can help with this!

GHENT UNIVERSITY

language and translation technology team

Limitations and dangers

GHENT UNIVERSITY

Limitations:

Historical🏛️ language remains difficult ⚠️

- less resourced
- spelling differences ✍🏻
- out-of-vocabulary📕
- stylistic differences:
  - more creative, metaphorical, poetic, lyrical 🎨

GHENT UNIVERSITY

language and translation technology team

Limitations

🛶 data drift:

📽️projecting modern-day cultural assumptions

🇪🇺🏰regional bias:
focus on Western world

⚠️⚠️TRIGGER WARNING⚠️⚠️

Limitations

Bias ⚖️🚨

- ethnic groups / slurs as fauna ☠️

Limitations

⚠️⚠️TRIGGER WARNING⚠️⚠️

Bias ⚖️🚨

- ethnic groups / slurs as fauna ☠️

Text	Entities
*[…] he with much difficulty prevailed on part of the **Indians** to begin some new plantation , that they might supply themselves with **grain**.*	**[part of the Indians]**
*Het , in het oor eens **Kaffers** allezins aangenaam luidend , gebulk van eene **koe** kan hem dermate verrukken […]*	[koe, **kaffers**]
*[…] hetzij door dieren of ook wel door den mensch — de **n*gers** gebruiken de wol daarvan als tonder — sterft de plant niet noodzakelijk […]*	[dieren, **n*gers***, plant]

GHENT UNIVERSITY

language and translation technology team

Limitations

Bias ⚖️🚨

- ethnic groups / slurs as fauna ☠️
- translation (anglophone bias)
- more West-European

Text		Entities
CHIEN** ET **LOUP	❯	[dog, wolf]
*zo is in Rusland de algemeene gewoonte van **Salmen** en andere soorten van **Visschen***	❯	[salmon,Visschen]

Ready to work?

GHENT
UNIVERSITY

What will we do?

📦 No annotated data ⇒ zero-shot

🧭great for exploratory work

💸 a bit of human-labeled data ⇒ few-shot

✨when you can provide some examples

👑💰a corpus of annotated data ⇒ training / fine-tuning

🎯specialized models, good at specific tasks

QR CODE

— GITHUB —