From Strings to Knowledge Graphs:

What Text2KG Still Gets Wrong

Raphael Troncy

INSTITUT CARNOT Télécom & Société numérique

EURECOM
Sophia Antipolis

Introducing the Knowledge Graph: things, not strings

Posted: Wednesday, May 16, 2012 g+1 7.7k Tweet 3,160 J'aime 3

Cross-posted on the Inside Search Blog

Search is a lot about discovery—the basic human need to learn and broaden your horizons. But searching still requires a lot of hard work by you, the user. So today I'm really excited to launch the Knowledge Graph, which will help you discover new information quickly and easily.

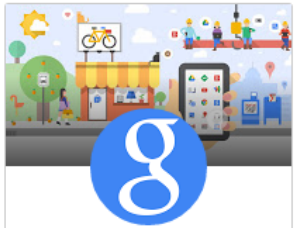
Take a query like [taj mahal]. For more than four decades, search has essentially been about matching keywords to queries. To a search engine the words [taj mahal] have been just that—two words.

But we all know that [taj mahal] has a much richer meaning. You might think of one of the world's most beautiful monuments, or a Grammy Award-winning musician, or possibly even a casino in Atlantic City, NJ. Or, depending on when you last ate, the nearest Indian restaurant. It's why we've been working on an intelligent model—in geek-speak, a "graph"—that understands real-world entities and their relationships to one another: things, not strings.

The Knowledge Graph enables you to search for things, people or places that Google knows about—landmarks, celebrities, cities, sports teams, buildings, geographical features, movies, celestial objects, works of art and more—and instantly get information that's relevant to your query. This is a critical first step towards building the next generation of search, which taps into the collective intelligence of the web and understands the world a bit more like people do.

https://blog.google/products/search/introducing-knowledge-graph-things-not/

depth. It currently contains more than 500 million objects, as well as more than 3.5 billion facts about and relationships between these different objects. And it's tuned based on what people search for, and what we find



Google

google.com/+google

News and updates on Google's products, technology and more

Follow +1

+11,105,298

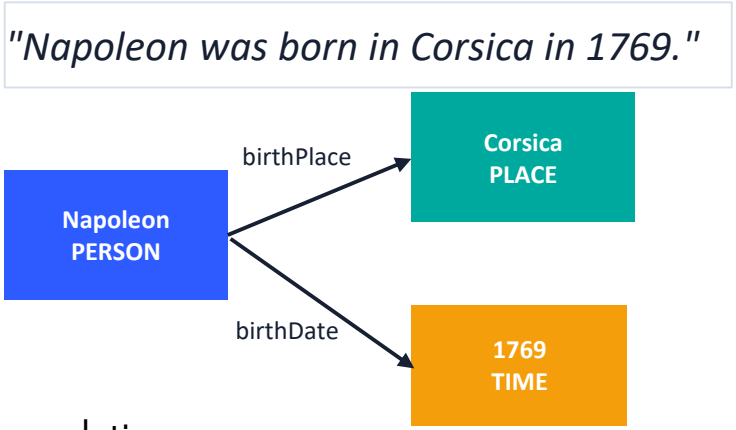
Labels

Archive

Feed

From Strings to Things … From Things to Graphs

- **Extracting triples 101**
  - Extracting named entities
  - Extracting relations



- **Text2KG is more than NER + relation extraction**
  - mention detection, entity typing, entity linking, co-reference resolution
  - attributes and relations extraction, frames, events, causal relationships
  - ontology alignment
  - validation (shapes), reasoning (consistency) and provenance

Open Information Extraction vs Closed Information Extraction

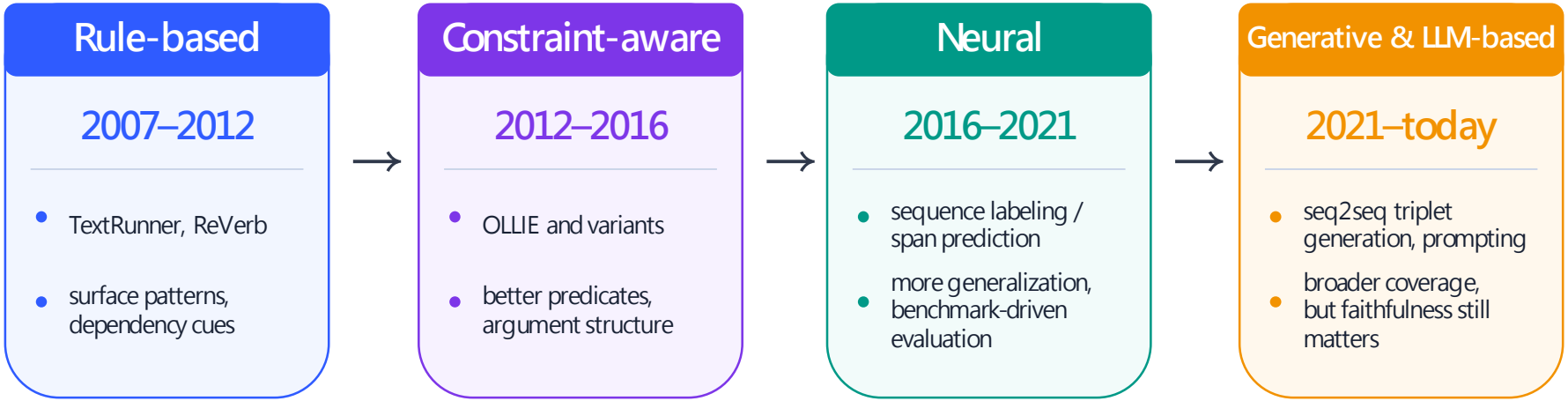
	Open IE	Ontology-guided IE
◎ **Goal**	Discover relational facts in text	Populate a graph that conforms to a conceptual model
▤ **Output**	Surface-oriented tuples / flexible predicates	Typed entities, relations, events, roles, constraints
✓ **Strength**	Fast discovery, broad coverage, low upfront modeling cost	Interoperability, reasoning, validation, reuse
△ **Typical weakness**	Heterogeneous predicates, redundancy, difficult integration	Higher annotation and modeling effort, narrower scope
☆ **Best use**	Exploration, corpus navigation, GraphRAG indexing	Knowledge engineering, domain applications, long-term assets

Open IE maximizes flexibility
Ontology-guided IE maximizes semantic precision and reuse

EURECOM Sophia Antipolis

OpenIE evolved with the modeling paradigm

The field moved from rule-based systems to neural models and now LLM-based extraction.



The task also evolved: from extracting surface tuples to supporting broader, more flexible open extraction settings.

EURECOM Sophia Antipolis

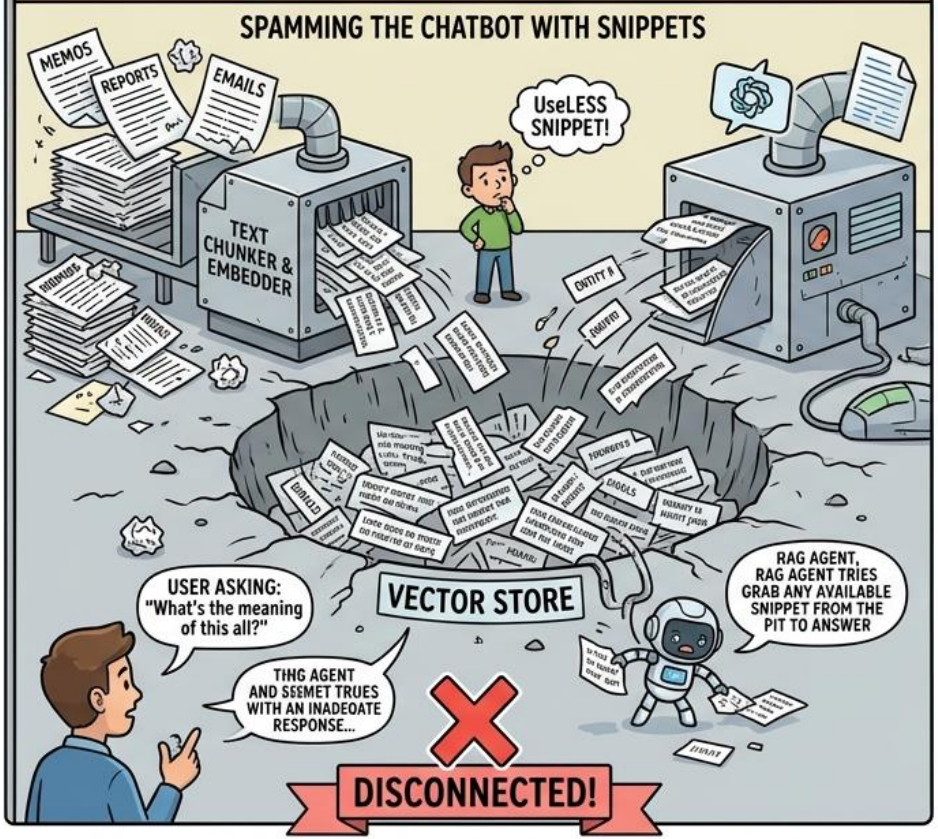
The GraphRAG Moment

EURECOM
Sophia Antipolis

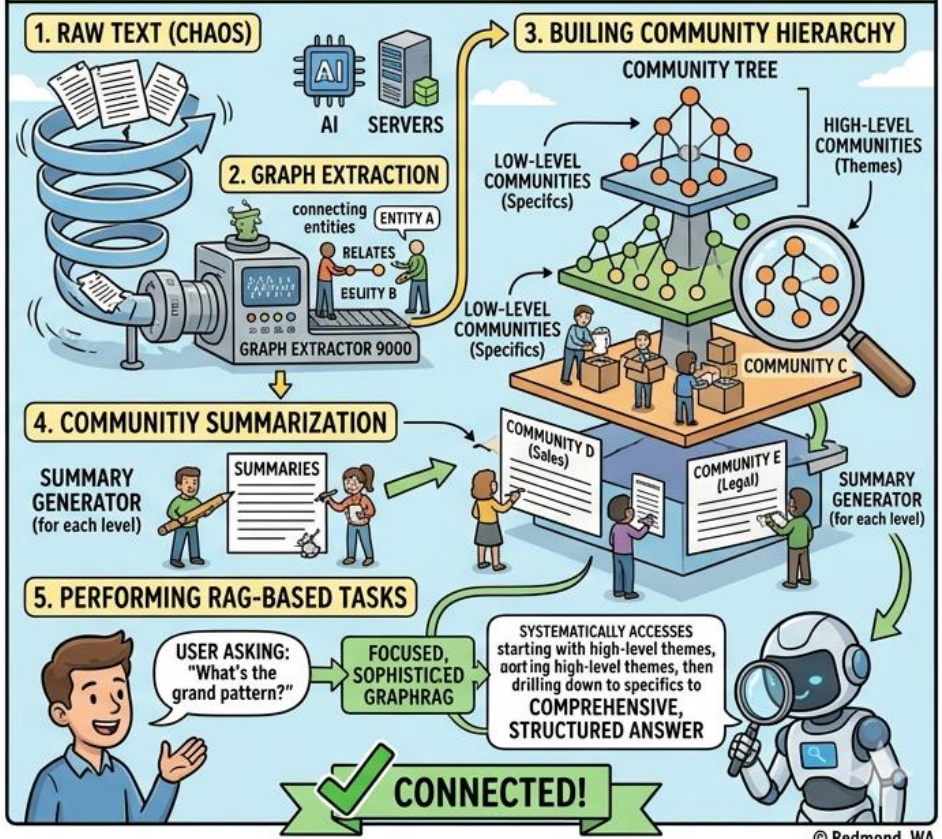
MICROSOFT GRAPHRAG: A STRUCTURED, HIERARCHICAL APPROACH TO RAG (VS. NAIVE SEMANTIC SEARCH)

REDMOND, WA

THE "NAIVE" METHOD: SEMANTIC SEARCH & CHUNKING



THE "SUPERIOR" GRAPHRAG METHOD: A STRUCTURED JOURNEY



© Redmond, WA

What Standard RAG misses

Chunk retrieval is good at local evidence lookup; GraphRAG targets global, thematic, and cross-document questions

Standard RAG	
Typical pipeline	Chunk → embed → retrieve top-k chunks → answer
Strong for	Specific questions, evidence lookup, citation-oriented QA
Weak for	“What are the main themes?”, “What tensions recur?”, “How does the corpus frame X?”
Failure mode	Retrieved chunks are individually relevant but do not provide a corpus-level synthesis

Great at finding evidence
Less good at “seeing the whole corpus”

GraphRAG	
Typical pipeline	Extract entities/relations → build graph → detect communities → summarize communities → answer
Strong for	Global questions, thematic overview, narrative synthesis, exploratory analysis
Key idea	Use graph structure and community summaries as query-time context, not only retrieved chunks
Trade-off	More preprocessing, more LLM-generated structure, and more uncertainty about semantic durability

Better for sensemaking
Not automatically a reusable knowledge graph

Why MSFT GraphRAG is useful – and why it is not yet a durable KG

Important distinction: a graph that helps answer questions vs a graph that can be validated, integrated, and reused.

What GraphRAG gets right

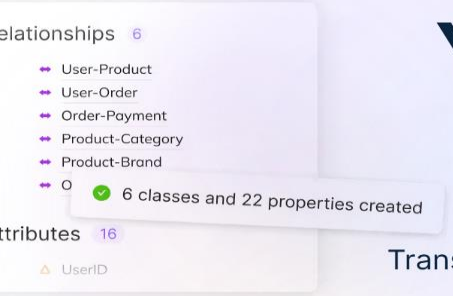
- It acknowledges that corpus-level questions are not simple retrieval tasks.
- It introduces useful intermediate structure: entities, relations, communities, summaries.
- It improves exploration, synthesis, and query-focused summarization over large corpora.
- It gives practitioners a practical graph layer without requiring a full ontology engineering effort.

What Text2KG should still criticize

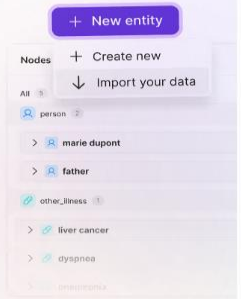
- The extracted graph is often an LLM-built index, not a semantically committed ontology-based KG.
- Predicates and entity types may remain loose, underspecified, or difficult to validate.
- Community summaries are valuable for answering questions, but they are not first-class knowledge assets.
- Interoperability, provenance, consistency checking, and long-term reuse remain limited.

Take-away: GraphRAG is a major step for retrieval and summarization —
but from a Text2KG perspective, a graph index is not automatically a durable knowledge graph

EURECOM Sophia Antipolis



Your documents deserve RAG done right.



Transparent and verifiable, our business-friendly Knowledge Studio makes sure AI reaches your enterprise-level needs.

Get started with GraphRAG

https://www.lettria.com/features/graphrag

If your business has high standards for querying documents, charts and tables...

Beyond Triples: extracting causal relationships

EURECOM
Sophia Antipolis

Motivation

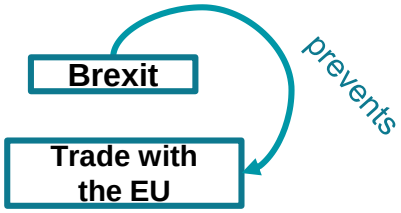
- Understanding **relations between events** is key for NLP
- But current approaches:
  - Focus on **simple relations** (temporal, coarse causality)
  - Struggle with **fine-grained reasoning**

✕ **No high-quality datasets** for fine-grained event relations

✕ Existing data is **limited, inconsistent, or too coarse**

Goal

Extracts fine-grained event relations from text

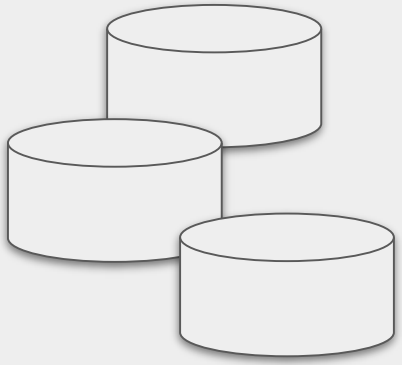


Contributions

- Dataset:
  - 500K+ sentences
  - 5 fine-grained event relations
- Model for Event Relation Extraction
- Comprehensive Evaluation
  - Pre-Trained Language Models
  - Seq2Seq models
  - LLMs

CausalSense

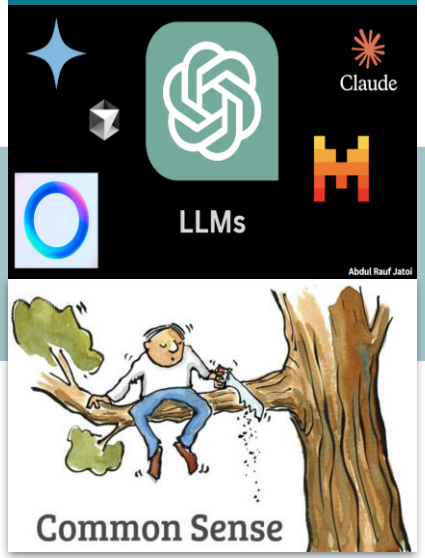
Mapping / Re-annotation



Existing Datasets

Rebboud et al. (2023)

Data Augmentation

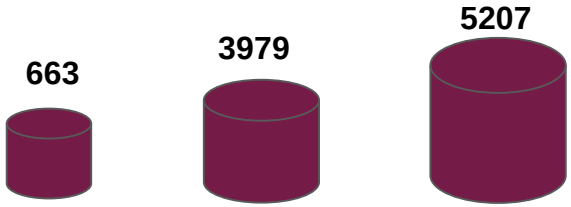


CausalSense

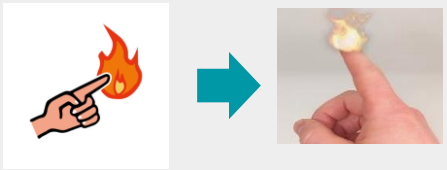
News Dataset

Category	Dataset	Total	Cause	Enable	Prevent	Intend	No-relation
News Data	Rebboud et al. (2023)	663	268	100	81	42	172
	Synthetic Data	1,228	0	350	419	459	0
	CNC	3,316	1,710	0	0	0	1,606
	Total News	5,207	1,978	450	500	501	1,778

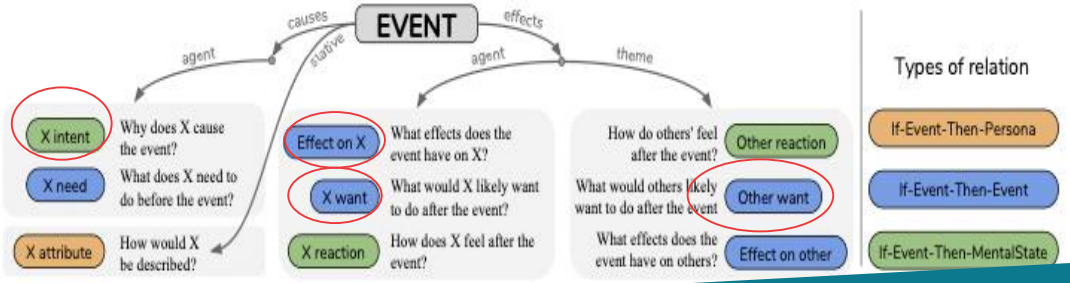
Data Augmentation Via Common Sense



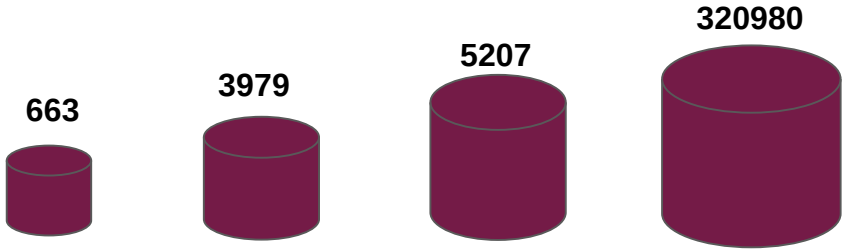
Common sense is the ability to infer likely causes and predict plausible effects based on a single observed event, using **everyday knowledge and reasoning**
(Sap et al., *ATOMIC,* 2019)



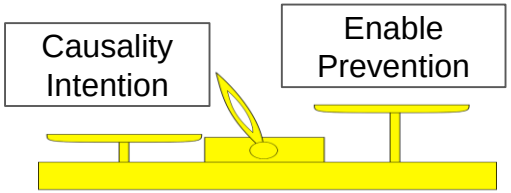
ATOMIC
An Atlas of Machine Commonsense for If-Then Reasoning



Data Augmentation Via Common Sense



Category	Dataset	Total	Cause	Enable	Prevent	Intend	No-relation
Common Sense	**ATOMIC**	315,173	82,242	0	0	146,588	86,943



Intention ☑
Causality ☑
Enable ☹
Prevention ☹

Data Augmentation Via Synthetic Commonsense

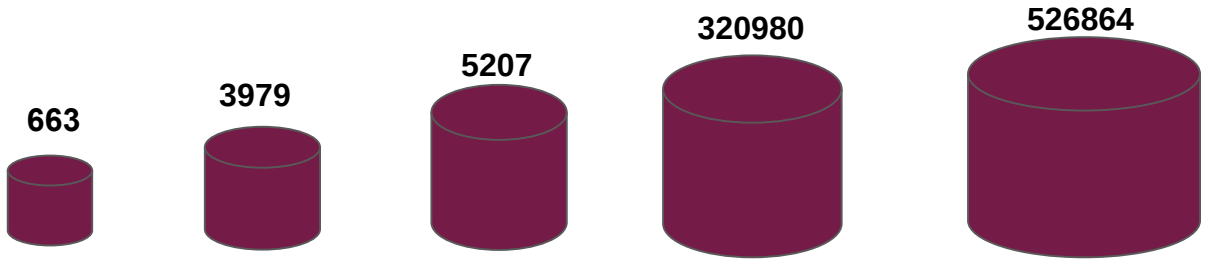
- Generate **synthetic** common sense to overcome data **imbalance**.
- Request the model to generate **common sense example** for the aimed relation
  - **Not guiding** with **example** for domain **diversity**

Request the LLM to generate few examples of the wanted event relation

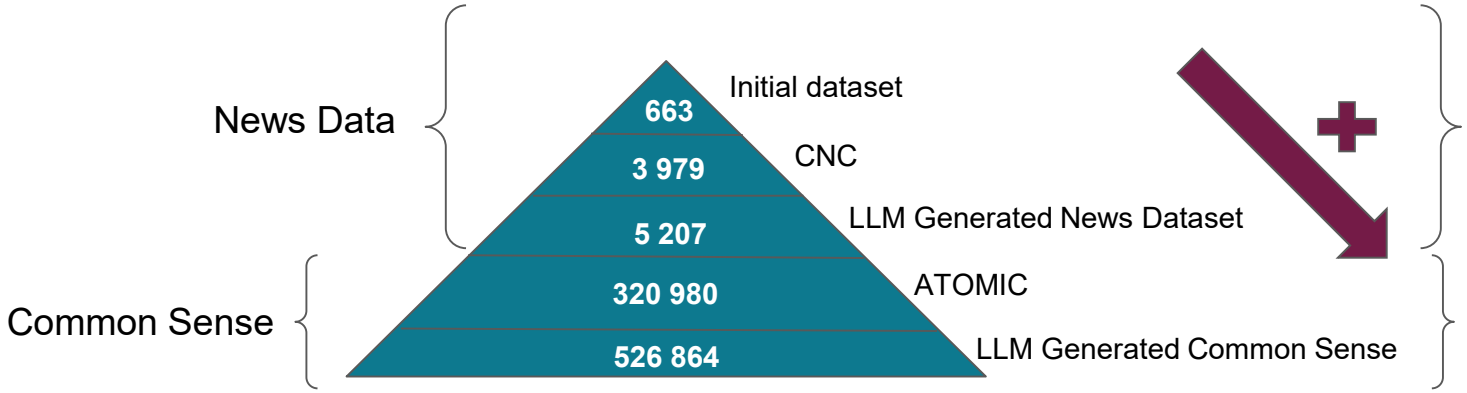
Give these examples as example seeds to the LLM

Manually validate these generated examples (100)

Data Augmentation Via Synthetic Commonsense



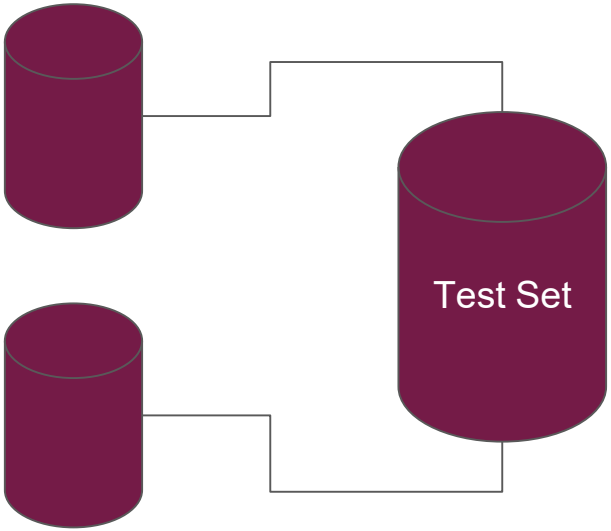
Category	Dataset	Total	Cause	Enable	Prevent	Intend	No-relation
Common Sense	ATOMIC	315,773	82,242	0	0	146,588	86,943
	Synth. Common Sense	205,884	0	65,485	53,456	0	86,943
	Total	521,657	82,242	65,485	53,456	146,588	173,886



Category	Total	Cause	Enable	Prevent	Intend	No-relation
News Data	5,207	1,978	450	500	501	1778
Common Sense	521,657	82,242	65,485	53,456	146,588	173,886
TOTAL Full Dataset	526,864	84,220	65,935	53,956	147,089	175,664

News Data

- **Non-synthetic** portion
- **Sample** from different relation **types**



AveriTec Dataset

AVeriTeC:

- 4,568 **real-world** claims with fact-check answers from 50 sources (e.g., CNN, Facebook)
- **Manual evaluation** of **4,130** samples

	Total	Cause	Enable	Prevent	Intend	No-relation
Test Set	632	351	89	52	40	100

Event Relation Extraction

3 sub-tasks

Relation Detection (RD)

- **Binary classification**
- Determine whether a **causal relation exists** in the sentence or not

Relation Classification (RC)

- **Sequence classification**
- Assign a sentence to target labels:
  - **Cause**
  - **Enable**
  - **Prevent**
  - **Intend**
  - **No-relation**

Event Extraction (EE)

- **Span detection**
- Precisely identify the subject and object segments (**event1** and **event2**)

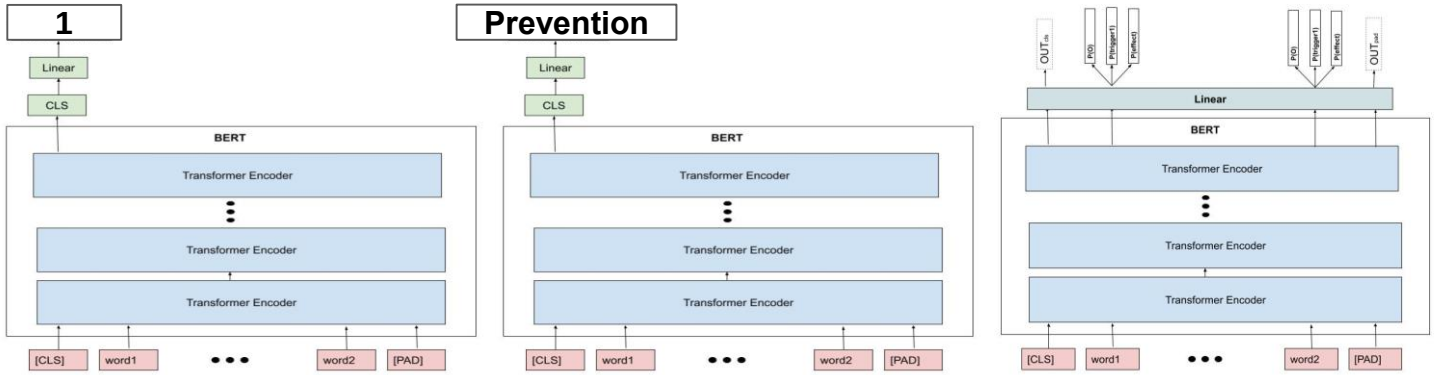
ERE

Separate Tasks

Multi-task Learning

LLMs

One Model for **each Task** *(trained and tested **separately**)*

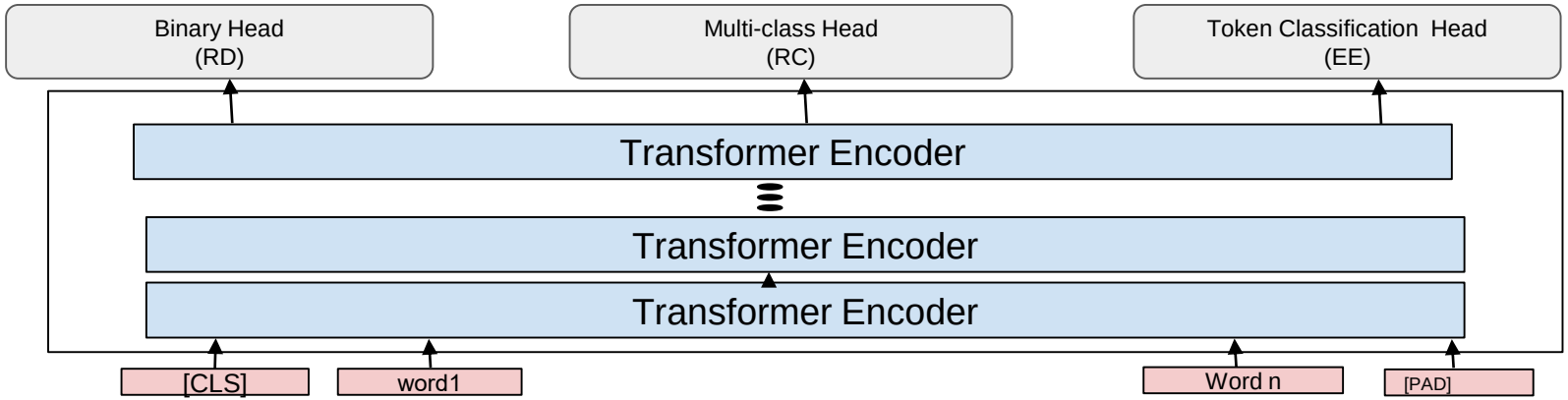


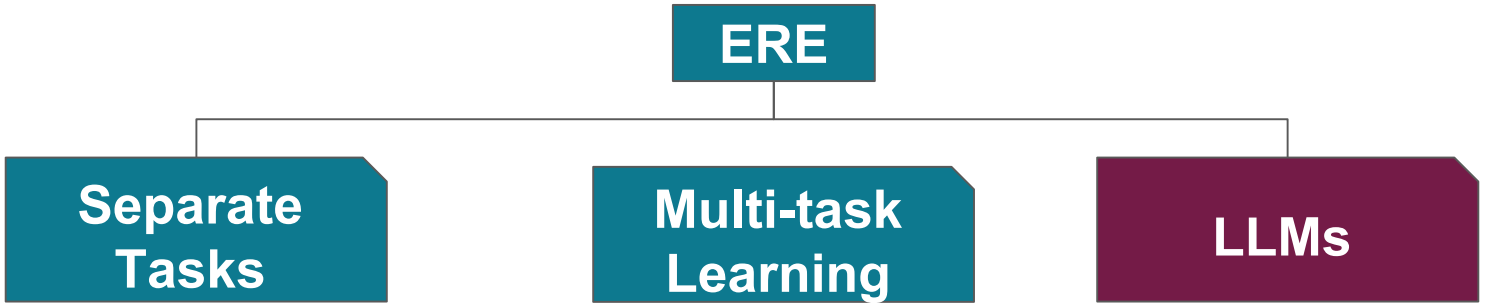
"One year after Brexit, the city is recruiting [...] trades with the EU in slow motion."

ERE

- **Separate Tasks**
- **Multi-task Learning**
- **LLMs**

3 heads trained **jointly**





- **Zero-shot** and **few-shot prompting**
- Included relation **definitions**
- Closed and open weights models (**GPT4** and **Zephyr**)

Results: Pre-trained Language Models

Train	Strategy	Model	RD F1	RC F1	EE F1	Avg F1
News	End-to-End	REBEL	0.56	0.65	0.59	0.60
	End-to-End	RoBERTa	0.98	0.74	0.20	0.64
	Separate	BERT	0.89	0.77	0.61	0.76
	Separate	RoBERTa	0.89	0.73	0.66	0.76
Combined (50% news, 50% common sense)	End-to-End	REBEL	0.60	0.75	**0.70**	0.68
	End-to-End	RoBERTa	**0.98**	**0.78**	0.20	0.65
	Separate	BERT	0.92	0.70	0.60	0.74
	Separate	RoBERTa	0.92	0.73	0.64	**0.763**

- **Relation Detection** performs well
- **Relation Classification** is harder macro F1: **0.78**
- **Enable** vs. **Cause**:
  Fine semantic line; often confused.
- **Prevent** misclassified as **Cause**
  <ARG1>The police have posted men in front of the office</ARG1>[...] <ARG0>to prevent any retaliatory attack by RSS men</ARG0>

REBEL: Relation Extraction By End-to-end Language generation (Huguet Cabot & Navigli, Findings 2021)

Results (LLMs)

Dataset	Strategy	Model	RD F1	RC F1	EE F1	Avg F1
Best PLM		RoBERTa	**0.92**	0.73	0.64	**0.763**
News	LLM	GPT-4 (0-shot)	0.29	0.33	0.23	0.29
	LLM	GPT-4 (2-shot)	**0.59**	0.53	0.42	0.51
	LLM	GPT-4 (4-shot)	0.57	**0.54**	**0.45**	**0.52**
Combined (50% news, 50% common sense)	LLM	Zephyr (4-shot)	0.29	0.10	0.20	0.19
	LLM	GPT-4 (2-shot)	0.49	0.46	0.35	0.43

- **GPT-4** in f**ew shots** shows better performance
- **Prompting** with Example **only with news better** than combining with common sense
- LLMs still l**ag behind** PLMs for ERE.

Our Contribution:

- Introduced **CausalSense**: 500K+ sentences)
- Leveraged **LLMs and common sense** for data augmentation and class balancing
- A model for **joint event extraction + relation classification**

Main Findings:

- **+32.3%** F1 improvement over state-of-the-art
- **End-to-end models** outperform pipeline approaches
- **Commonsense** significantly boosts performance

CODE AND DATA

bit.ly/kflow-rel-extraction

LREC 2026 Palma

Beyond Triples: extracting frame semantics in the Cultural Heritage sector

Smell experiences: ODEUROPA
Textile: SILKNOW

EURECOM
Sophia Antipolis

Centuries of collective smell memories

*What are the **most frequent** smell sources in **London** in the 18th century?*

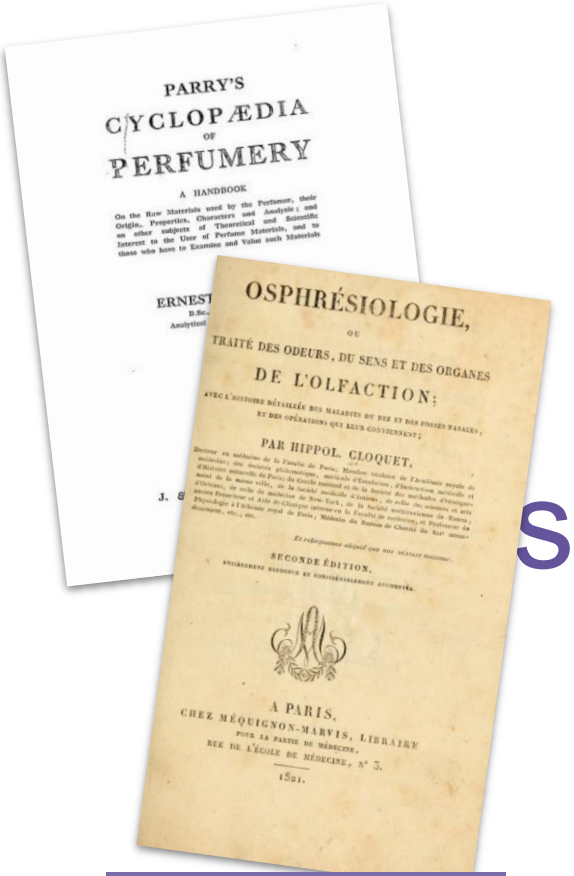
*When did the smell of **pollution** start to be mentioned?*

*What smells were perceived during the **Waterloo Battle**?*

*What **emotions** were associated with floral smells in the 19th century?*

Heritage

*How have the **adjectives** used for describing a smell change over time ?*



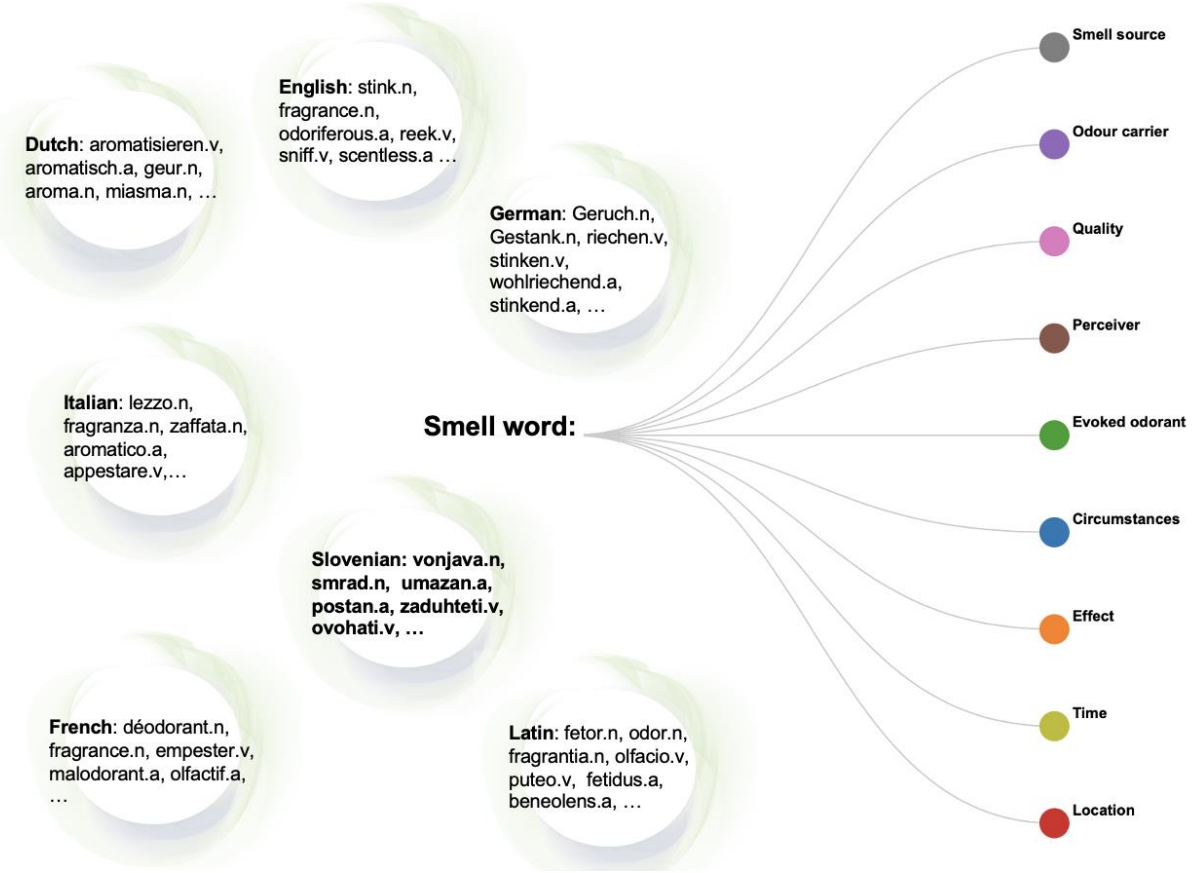
Centuries of collective smell memories

Heritage

Text

Images

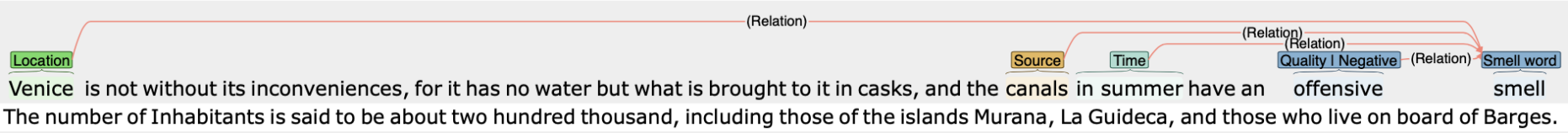
Extracting Olfactory Events from Multilingual Documents

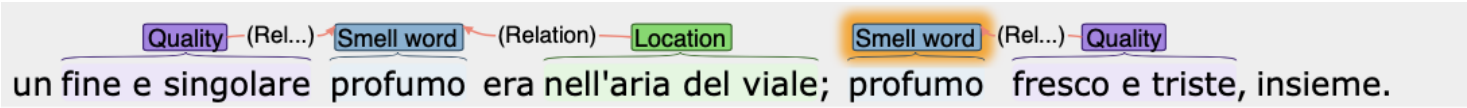


Annotation Scheme

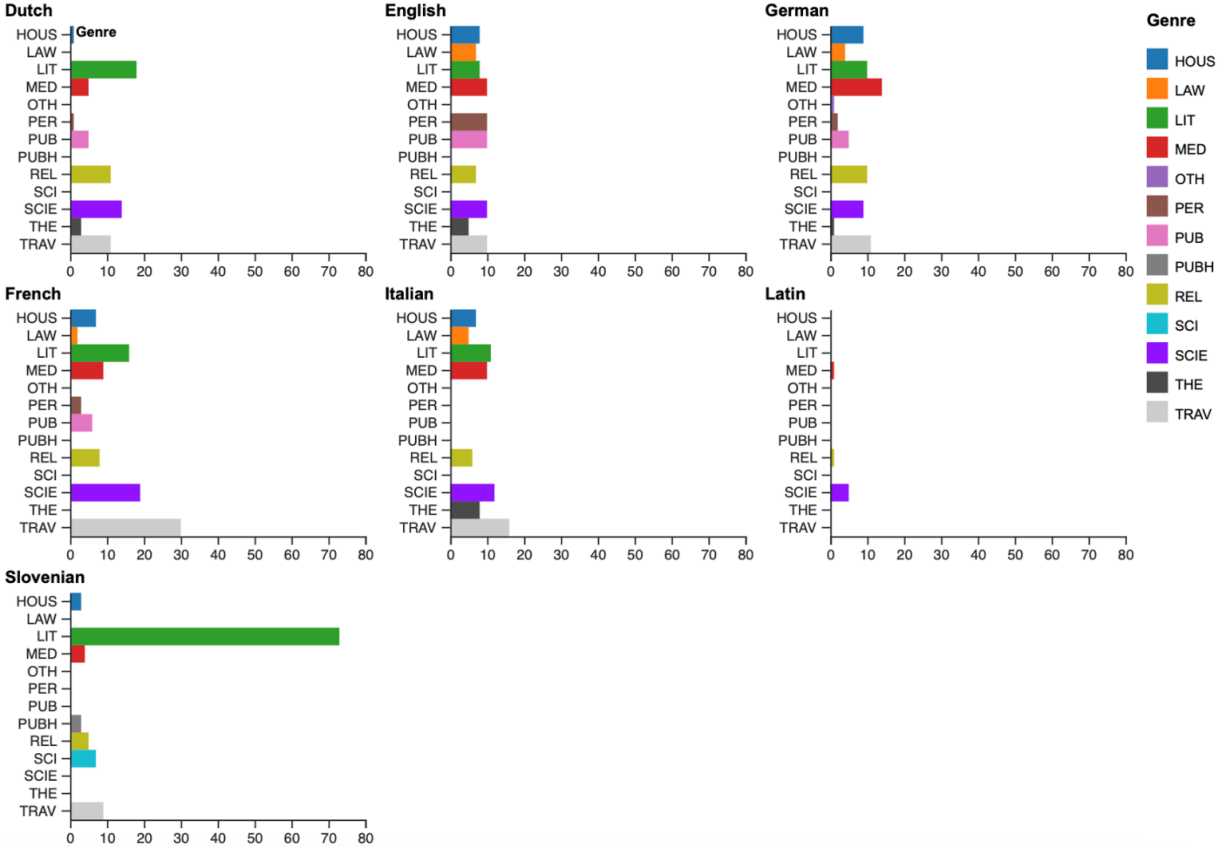
- **Odorant**: The person, object or place that has a specific smell.
- **Quality**: Quality associated with a smell. For example *rancid*, *fresh*, etc. This is typically expressed by qualitative adjectives.
- **Perceiver**: The entity that perceives an odour, the being who has a perceptual experience, not necessarily on purpose.
- **Evoked odorant**: The object, place or similar that is evoked by the odour, even if it is not visible in the scene.
- **Location**: The place where the olfactory situation or event takes place
- **Circumstances**: The circumstances under which the olfactory situation or event takes place
- **Time**: Temporal expressions, including frequency and duration, that characterise the olfactory situation

Annotation Interface [INCEpTION]

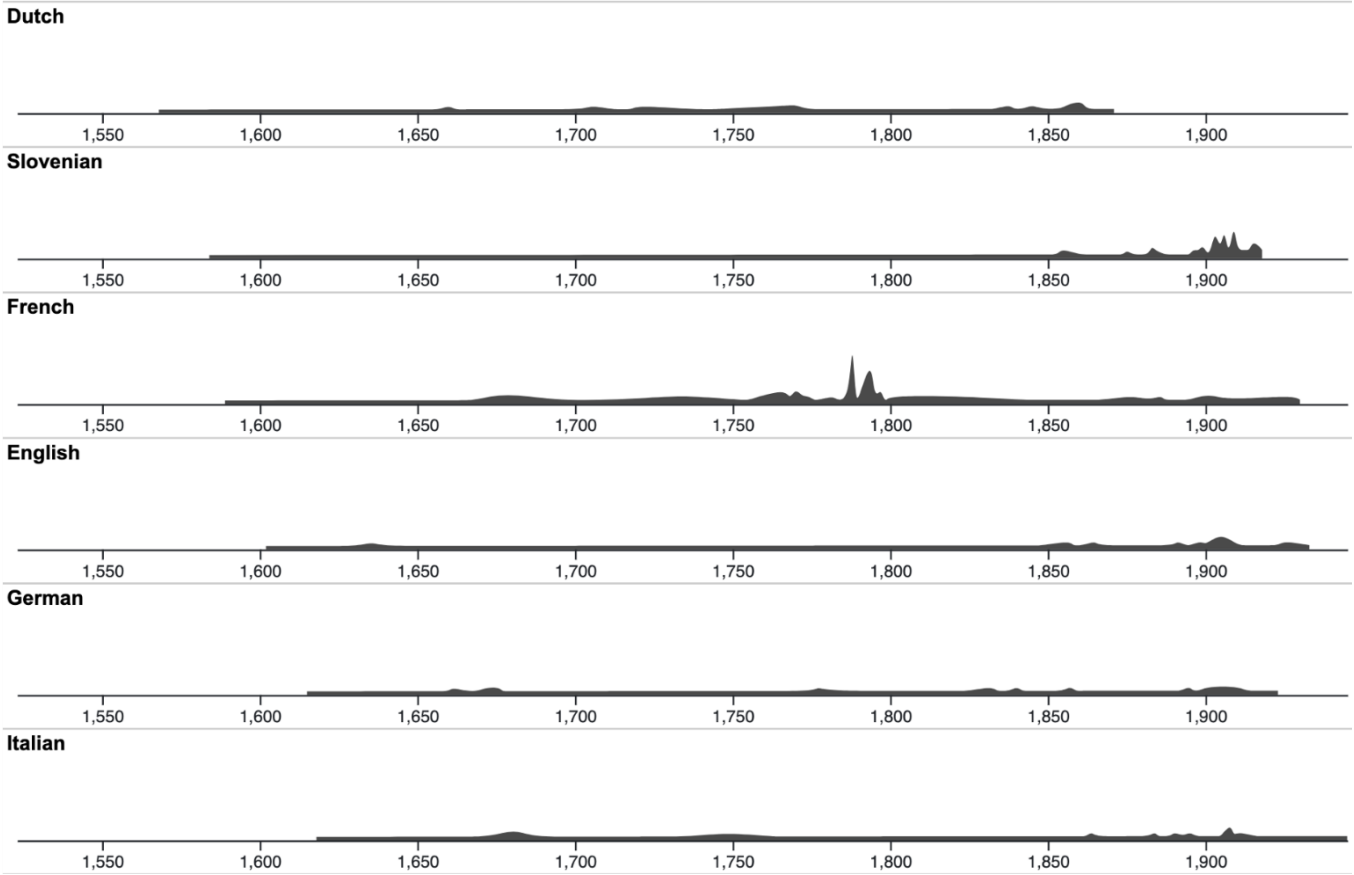




Benchmark genre distribution



Benchmark temporal distribution



Benchmark content

	Dutch	English	French	German	Italian	Latin	Slovene
Smell events	1,929	1,530	664	1,493	1,228	1,199	1,1917

Top event patterns in all languages are:

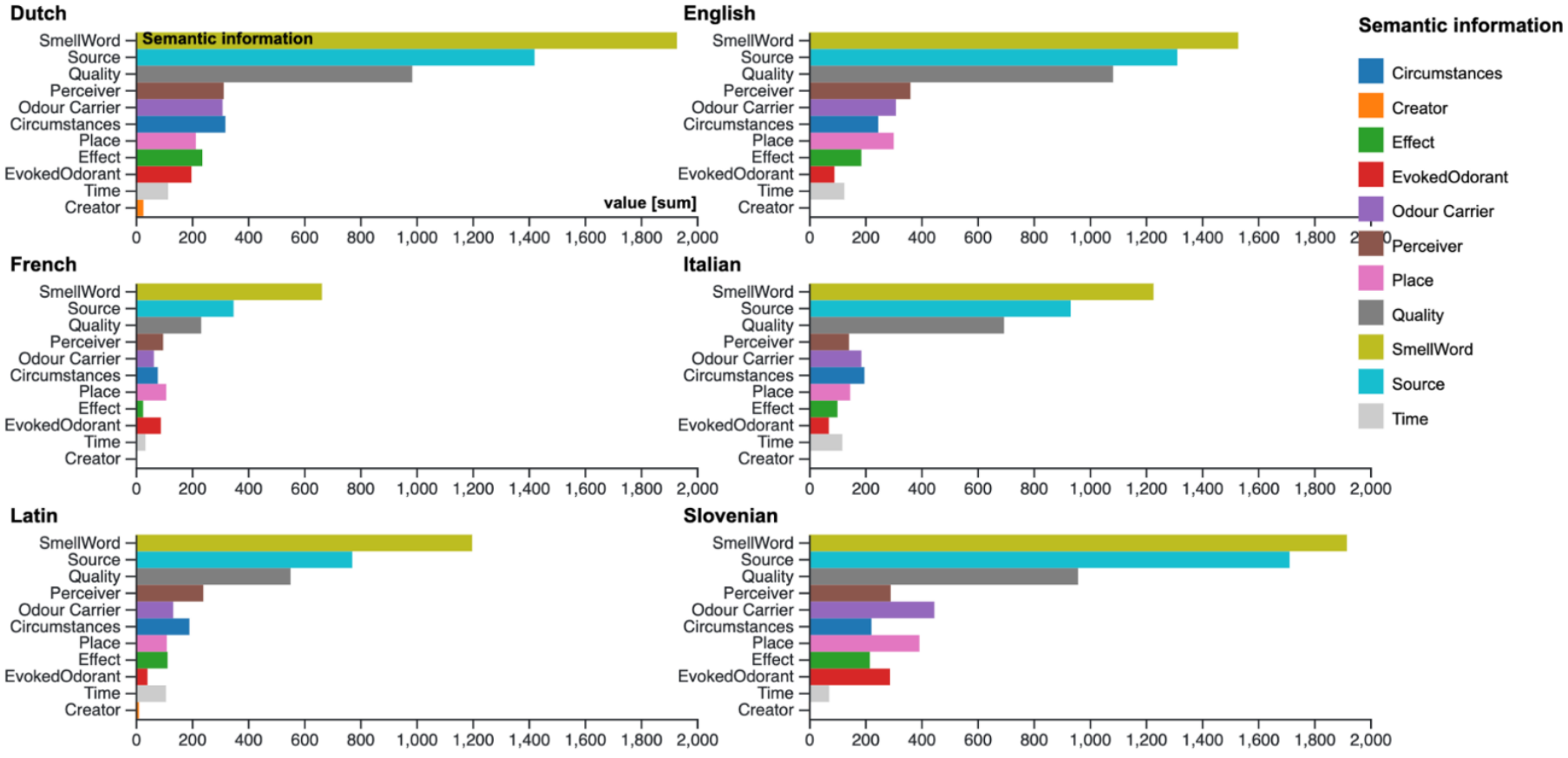
Smell word + **[Smell source]** + **[Quality]**

Smell word + **[Smell source]**

Smell word + **[Quality]**

Smell word

Benchmark content



Odeuropa text processing system

English - Multitask. Average F1 on 10 folds

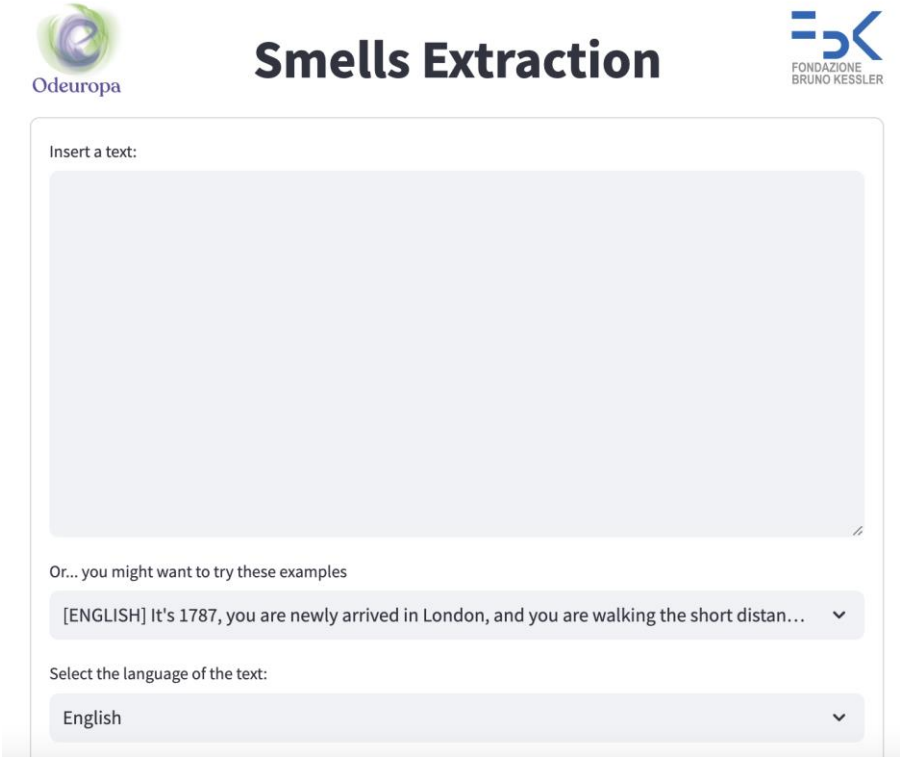
Model	Training Metric	Smell Word	Smell Source	Quality	Odour Carrier	Evoked Odorant	Location	Perceiver	Time	Circumstances	Effect
Monolingual training data											
monolingual	span-f1	**0,871**	0,571	**0,758**	**0,482**	**0,572**	**0,542**	**0,510**	0,434	0,461	**0,405**
monolingual	token-f1	0,864	0,571	0,759	0,483	0,535	0,535	0,484	0,417	0,480	0,365
multilingual	span-f1	0,860	**0,583**	0,747	0,452	0,542	0,560	0,488	0,471	0,455	0,271
multilingual	token-f1	0,860	0,533	0,741	0,441	0,473	0,500	0,461	0,440	0,457	0,296
Multilingual training data											
multilingual	span-f1	0,865	0,574	0,759	0,462	0,517	0,546	0,488	**0,528**	**0,480**	0,339
multilingual	token-f1	0,783	0,536	0,745	0,479	0,552	0,508	0,489	0,465	0,452	0,347

On the performance of historical Language Models

	Model	Quality	Smell Source	Smell Word	Overall
DE	bert-base-historical-german-rw-cased	35.76	17.64	75.46	43.94
	gbert-base	33.59	20.40	77.30	*47.71*
EN	MacBERTh	70.56	50.46	88.09	**69.85**
	bert-base-uncased	68.30	46.83	86.42	66.98
FR	D'alemBERT	45.30	31.87	81.79	54.26
	CamemBERT [33]	77.6	38.7	55.9	*57.4*
IT	BERToldo 1500-1700	67.93	39.63	86.12	69.17
	BERToldo 1700-1900	72.20	50.60	72.60	66.00
	BERToldo all	67.70	37.50	87.50	69.70
	bert-base-italian-uncased	68.26	38.97	87.12	*70.27*
NL	GysBERT	41.99	13.26	75.80	**49.11**
	BERTje	39.72	13.17	75.55	46.58

Table 3: F1 scores of the LLMs on the smell event frames Quality, Smell Source, and Smell Word. The column *Overall* is the F1-micro of the total extraction performance. The bold face indicates the settings LLM-Hs outperform LLM-Cs.

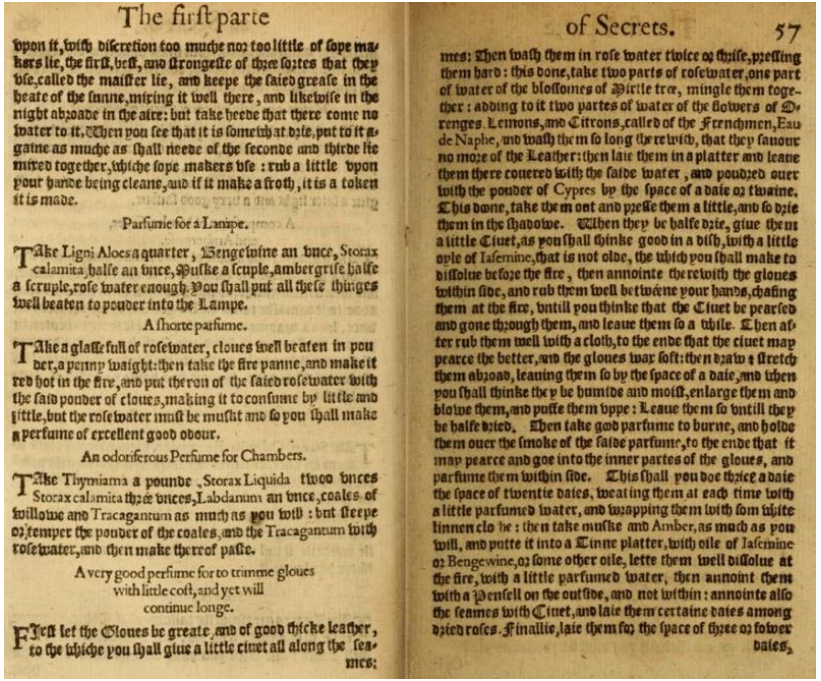
Demonstrator: Olfactory Information Extraction



Olfactory Information Extraction: Frame Semantics

[ENGLISH] It ' s 1787 , you are newly arrived in London , and you are walking the short distance from the Saracen ' s Head Inn to the nearby Newgate prison . As you pass | Circumstances the Old Bailey courthouse | Location you | Perceiver catch a terrible | Quality smell | Smell Word in the air | Odour Carrier . Uncertain of its origins , you ask a lawyer as they hurry past on their way to a trial . They tell you that the smell | Smell Word arose from the burning | Circumstances of a woman who had been found guilty of coining farthings | Smell Source . The public burning of women in England only ended in 1790 , Catherine Hayes being the last such individual to be thus punished . Up until 1789 | Time the scent | Smell Word of burnt flesh | Smell Source also appeared in the courtroom itself | Location , where some malefactors | Perceiver might be branded with a hot iron - " T " for theft , " F " for felon , or " M " for murder . The smell | Smell Word of burning | Smell Source was a warning to others | Effect . But smell | Smell Word could also feature as part of the humilitation of legal or , in some cases , extra - judicial punishment | Effect .

Verie good perfume for to trimme gloues with litle cost, and yet will continue longe



The first parte

vpon it, with discretion too muche nor too little of sope makers lie, the first, best, and strongeste of three sortes that they vse, called the maister lie, and keepe the saied grease in the heate of the sunne, mixing it well there, and likewise in the night abroade in the aire: but take heede that there come no water to it. When you see that it is somewhat dryed, put to it againe as muche as shall needs of the seconde and thirde lie mixed together, whiche sope makers vse: rub a little vpon your hande being cleane, and if it make a froth, it is a token it is made.

Parfume for a Lampe.

Take Ligni Aloes a quarter, Bengewine an vnce, Storax calamita, halfe an vnce, Muske a scruple, ambergrise halfe a scruple, rose water enough. You shall put all these thinges well beaten to powder into the Lampe.

A shorte parfume.

Take a glasse full of rosewater, cloues well beaten in powder, a penny waight: then take the fire panne, and make it red hot in the fire, and put theron of the saied rosewater with the said powder of cloues, making it to consume by little and little, but the rosewater must be muskt and so you shall make a perfume of excellent good odour.

An odoriferous Perfume for Chambers.

Take Thymiama a pounde, Storax Liquida two vnces Storax calamita three vnces, Labdanum an vnce, coales of willowe and Tracagantum as much as you will: but steepe or temper the powder of the coales, and the Tracagantum with rosewater, and then make thereof paste.

A very good perfume for to trimme gloues with little cost, and yet will continue longe.

First let the Gloues be greate, and of good thicke leather, to the whiche you shall giue a little ciuet all along the seames:

of Secrets. 57

mes: Then wash them in rose water twice or thrise, pressing them hard: this done, take two parts of rosewater, one part of water of the blossomes of Myrtle tree, mingle them together: adding to it two partes of water of the flowers of Orenges, Lemons, and Citrons, called of the Frenchmen, Eau de Naphe, and wash them so long therewith, that they sauour no more of the Leather: then laie them in a platter and leane them there couered with the saide water, and poudred ouer with the pouder of Cypres by the space of a daie or twaine. This done, take them out and presse them a little, and so drie them in the shadowe. When they be halfe drie, giue them a little Ciuet, as you shall thinke good in a dish, with a little oyle of Iasemine, that is not olde, the which you shall make to dissolue before the fire, then annointe therewith the gloues within side, and rub them well betweene your hands, chafing them at the fire, vntill you thinke that the Ciuet be pearsed and gone through them, and leaue them so a while. Then after rub them well with a cloth, to the ende that the ciuet may pearce the better, and the gloues may soft: then drawe & stretch them abroad, leauing them so by the space of a daie, and when you shall thinke they be humide and moist, enlarge them and blowe them, and putte them vppe: Leaue them so vntill they be halfe dried. Then take good parfume to burne, and holde them ouer the smoke of the saide parfume, to the ende that it may pearce and goe into the inner partes of the gloues, and parfume them within side. This shall you doe thrice a daie the space of twentie daies, weating them at each time with a little parfumed water, and wrapping them with som white linnen cloth: then take muske and Amber, as much as you will, and putte it into a Tinne platter, with oile of Iasemine or Bengewine, or some other oile, lette them well dissolue at the fire, with a little parfumed water, then annoint them with a Pensell on the outside, and not within: annointe also the seames with Ciuet, and laie them certaine daies among dried roses. Finallie, laie them for the space of three or fower daies,

Ingredients:
rose water; myrtle blossom water; orange, lemon, and citron water; perfumed water; cypress powder; jasmine oil; ben oil; perfume (*probably incense*); dried roses; civet; musk; and ambergris.

Girolamo Ruscelli, *The Secrets of the Reuerende Maister Alexis of Piemount*, trans. Wyllyam Warde (London, 1558)

FIrst let the gloues bee greate, and of good thicke leather, to the which you shall gyue a little **Ciuet** all alonge the seames: Than washe them in **rose water**, twise or thryse, pressing theym harde: this doen, take twoo partes of **rose water**, one parte of the **water of the blossoms of Mirtell** tree, mingle them together: addinge to it two partes of the **water of the flowres of Orenges**, **Lemons, & Citrons**, called of the Frēchmen, can de nafe, and washe them so long therwith, that they **sauour no moore of the leather** [...]

Than wil they bee **excellent**, as if it were to present an **emperour** withall.

-- Girolamo Ruscelli. The Secrets of the Reuerende Maister Alexis of Piemount. 1558

civet

rose water

dried rose

musk

amber

myrtle blossom water

orange flower water

Jasmine oil

hedonic

excellent

crm:P2 has type

crm:P141 assigned

crm:E13 Attribute Assignment

od:L12 Smell Emission

od:L11 Smell

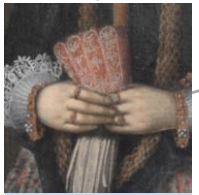
od:F2 has source

od:F2 has source

crm:P140 assigned attribute to

crm:P17 was motivated by

crm: P14 carried out by



gloves

od:F3 has carrier

od:L14 Smell Transformation

crm:P33 used specific technique

crm:E29 Design or Procedure

od:F1 generated

od:L11 Smell

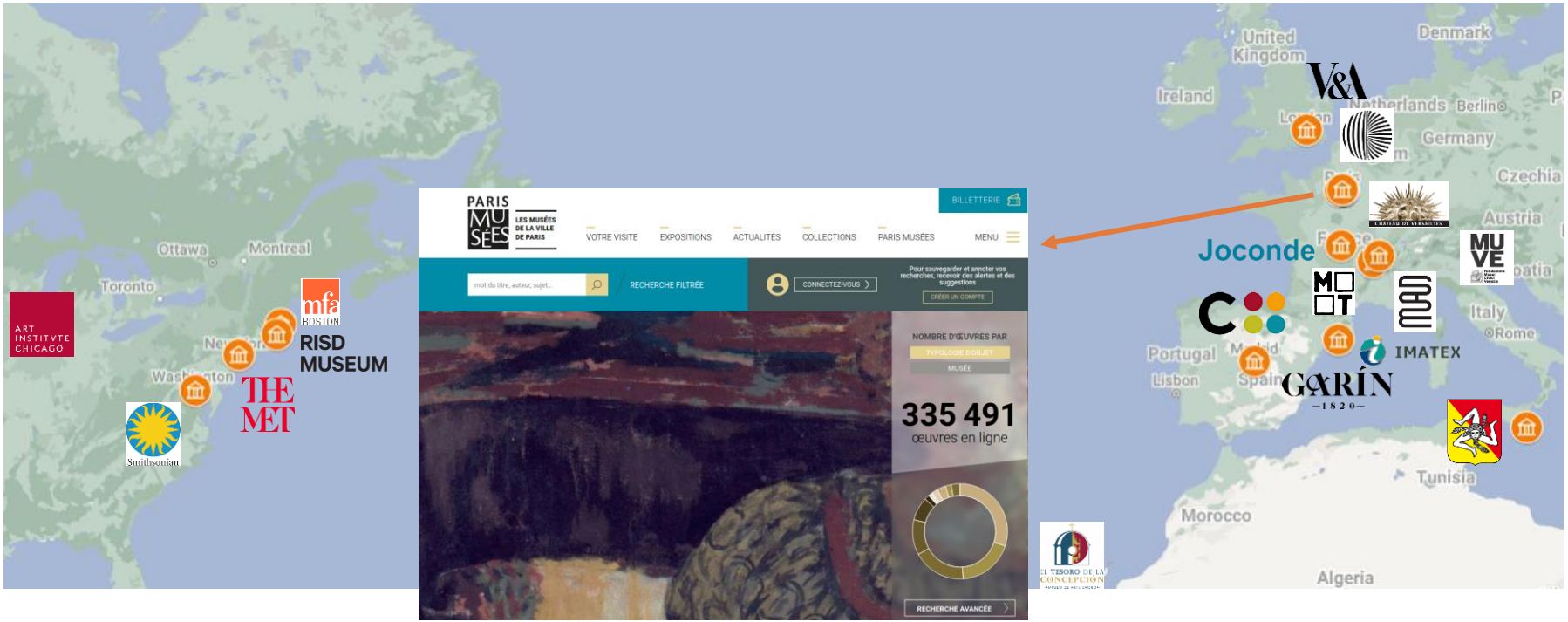
od:F2 perceived

od:L13 Olfactory Experience

Emperor

crm:E21 Person

Historians and museum textiles logbook



- Manually searching for candidate museums based on their silk collections
- Only select museums with relevant silk items and image illustrations

CDMT Terassa - 4537

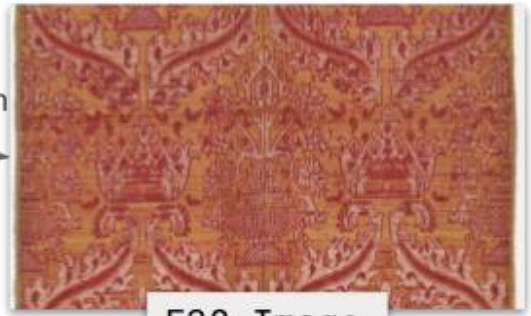
1576 / 1600

P86 falls within

16th century (dates CE)

E22 Man-Made Object

P138i has representation



E38 Image

Brocatelle@en
Brocatel@es
Brocatelle@fr
Broccatello@it

P4 has time span

P32 used general technique

P108 has produced

E12 Production

P41 classified

L4 assigned domain type

preferred Label

Textile@en
Tejido@es

Silk bombyx mori@en
Seda bombyx mori@es
Soie du Bombyx du mûrier@fr
Seta bombyx mori@it

P126 employed

P8 took place on or within

Italy

P65 shows visual item

Crown@en
Corona@es
Couronne@fr
Corona@it

MET - 37.80.1

sixteenth century
(dates CE)

P4 has
time span

E22 Man-Made
Object

P108 has
produced

P138i has
representation

E38 Image

O8 observed

S4 Observation

"This tapestry is one of seven hangings at The Cloisters that depict the hunt of the unicorn....."

P3 has
note

E12 Production

P32 used
general
technique

?

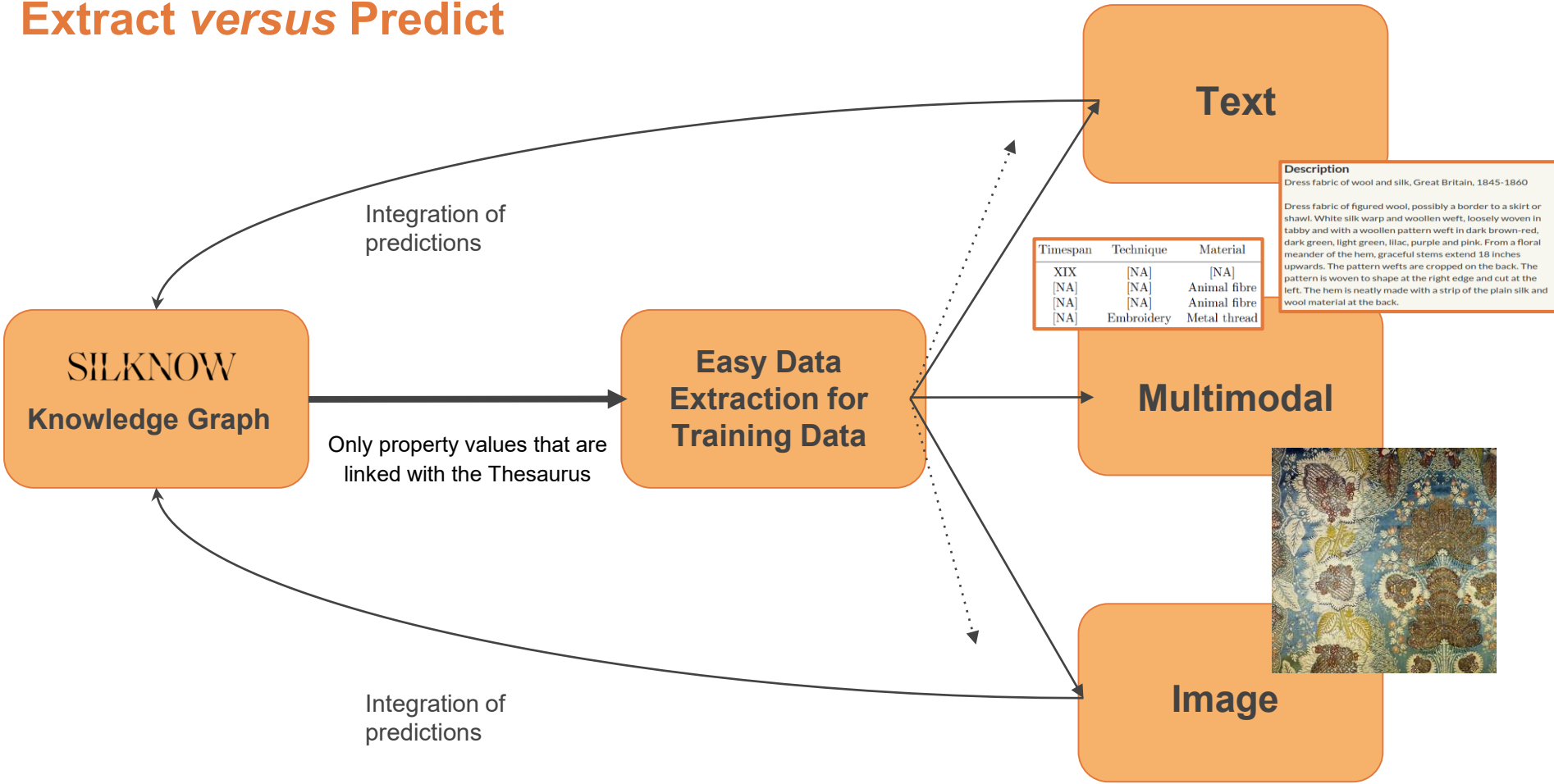
P126 employed

Wool@en
Lana@es
Laine@fr
Lana@it

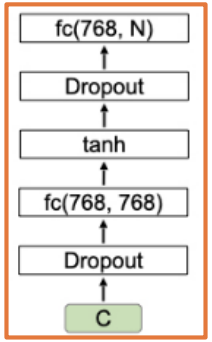
P8 took place on
or within

France

Extract *versus* Predict



Supervised Text Classification - Methodology



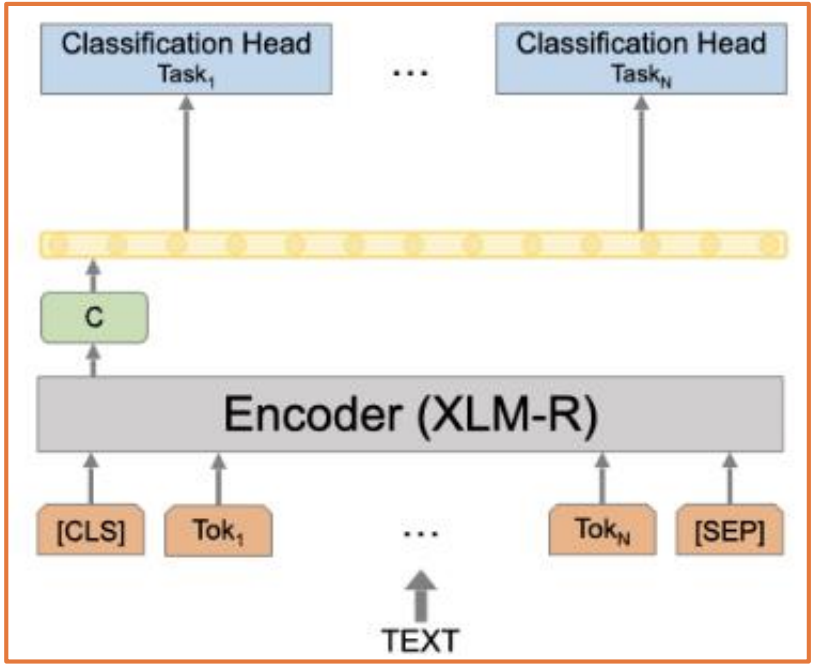
Training:

Examples of objects' text descriptions with the property it needs to learn to infer

Evaluation:

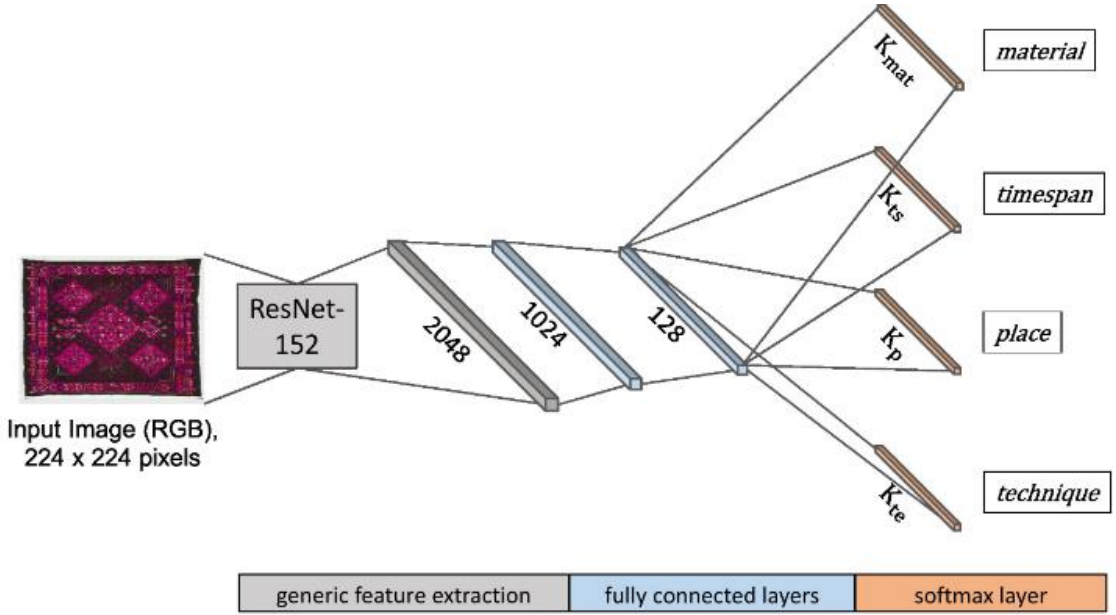
We feed it text but hold out the property value and see if the model guesses it correctly

Based on a shared fine-tuned **XLM-R encoder**, due to preliminary architecture comparisons and because it provides cross-linguality



Supervised Image Classification - Methodology

- Multi-task CNN (ResNet 152[*He et al., 2016*]), due to its proven success with image classification, pre-trained on ImageNet
- 4 output branches, each for one semantic property
- Training on examples with both an image and the semantic property it needs to learn

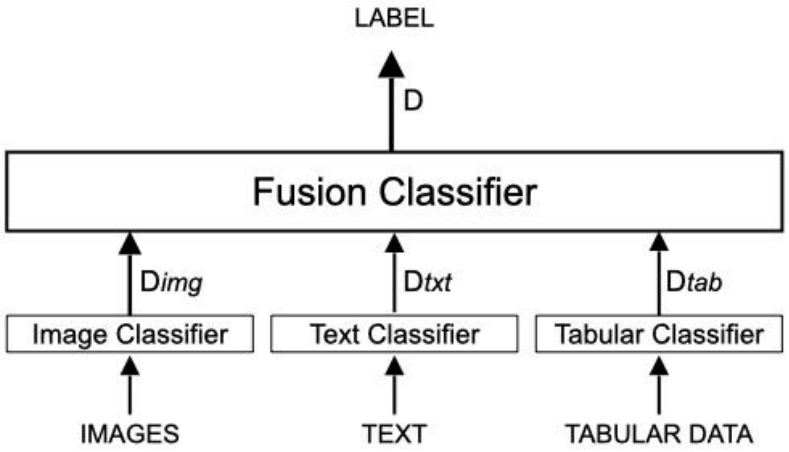


K. He, X. Zhang, S. Ren, and J. Sun. 2016. Identity mappings in deep residualnetworks. InEuropean Conference on Computer Vision (ECCV). 630–645.

Multimodal and tabular classification

- Gradient Boosted Decision Trees in both cases
- XGBoost implementation

Target Variable	Target Value	Feature				
		museum	place	timpespan	technique	material
place	FR	risd	-	[NA]	[NA]	animal fibre
timespan	XVIII	met	[NA]	-	embroidery	animal fibre
technique	other	garin	ES	XX	-	vegetal fibre
material	vegetable fibre	vam	GB	XIX	embroidery	-



Tabular classification input with one example per row task

Results

Variable	Nr. of Classes	Train Set 60%	Valid. Set 20%	Test Set 20%	Image	Text	Tabular	Multimodal
place	9	10,435	3,456	3,470	38.0	65.0	46.2	**77.6**
timespan	5	8,819	2,975	2,949	49.2	55.6	58.6	**74.2**
technique	4	4,813	1,663	1,675	73.5	41.0	68.3	**83.6**
material	3	12,865	4,263	4,351	46.5	37.4	49.4	**61.3**

Harmonized F1 scores in % for evaluation on the test set

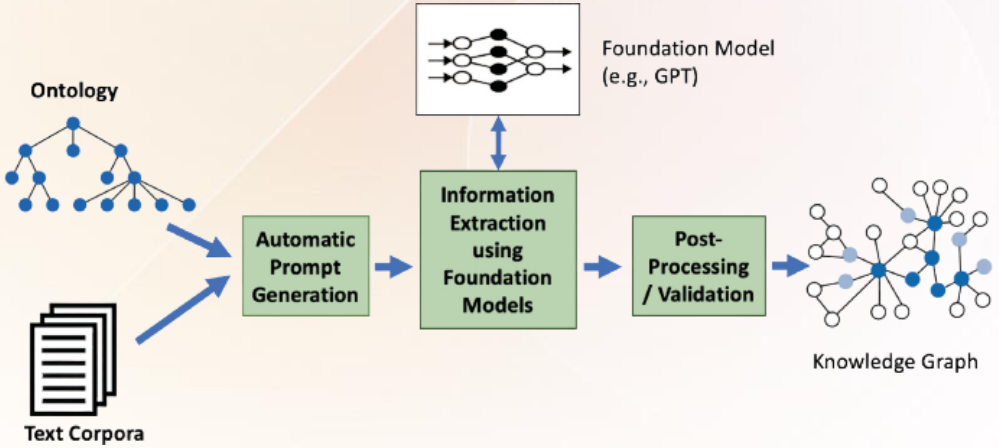
Benchmarking and Evaluation

EURECOM
Sophia Antipolis

State-of-art

Text2KGBench

- Presented at ISWC 2023
- Benchmark composed of
  - TekGen corpus
    13k+ sentences, 10 ontologies
  - WebNLG corpus Our focus!
    4860 sentences, 19 ontologies
- Repo: github.com/cenguix/Text2KGBench



Mihindukulasooriya, N., Tiwari, S., Enguix, C.F., Lata, K. (2023). **Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text**. In: The Semantic Web – ISWC 2023. Lecture Notes in Computer Science, vol 14266. Springer, Cham. https://doi.org/10.1007/978-3-031-47243-5_14

Problem

Critical Flaws in Text2KGBench

Ontological Flaws

✕ Flat Structure

No formal class hierarchy (rdfs:subClassOf).

✕ Semantic Ambiguity

Out-of-domain concepts and overly generic properties (e.g., location).

✕ Lack of Rigor

No formal distinction between object and datatype properties.

Annotation & Data Quality Flaws

✕ Inconsistency

No standardized format for entities or literal values (dates, numbers).

✕ Lack of Grounding

Annotations often relied on external knowledge.

Structural & Technical Flaws

✕ Poor Usability

Missing metadata, undocumented ontologies, and complex URIs

Our Solution

Text2KGBench-LettrIA

A substantially revised and curated benchmark

~1000 hours of human labor

Systematic Ontology Refinement

Revised 19 domain ontologies to enforce hierarchical structure (`subClassOf`) and formal typing.

Complete Re-Annotation

Re-annotated all 4,860 sentences, yielding over 14,000 high-fidelity, text-grounded triples.

Enriched Data Format

Introduced enhanced metadata and explicit typing to ensure reproducibility and support multifaceted evaluation.

Methodology

Ontology Refinement

Semantic Coherence	
Pruned irrelevant concepts	(removed `Club` from `Film` ontology)
Harmonized and specified property names	(`staff` →`academicStaff`)

Structural Enhancements	
Introduced a formal class hierarchy	(University ⊂ AcademicInstitution ⊂ Organization)
More nuanced evaluation	(rewarding a correct superclass prediction)

Formal Typing & Usability
Rigorously defined ObjectProperty and DatatypeProperty with explicit domains and ranges.
Added rdfs:comment annotations and simplified URIs.

Computing Diff

Ontology Statistics

Ontology Name	LettrIA			Text2KGBench	
	Classes	Object Prop.	Datatype Prop.	Classes	Object Prop.
airport	11	15	5	13	39
artist	12	16	7	19	39
astronaut	18	10	11	15	38
athlete	10	18	9	14	37
building	9	11	12	14	38
celestialbody	5	1	17	8	27
city	13	10	10	11	23
comicscharacter	8	8	4	10	18
company	9	13	6	10	28
film	5	10	5	18	44
food	12	13	2	12	24
meanoftransportation	12	20	28	20	68
monument	10	10	4	14	26
musicalwork	15	22	3	15	35
politician	17	25	9	19	40
scientist	12	15	5	15	47
sportsteam	9	12	3	14	24
university	11	16	11	15	46
writtenwork	10	17	13	10	44
TOTAL	**208**	**262**	**164**	**266**	**685**

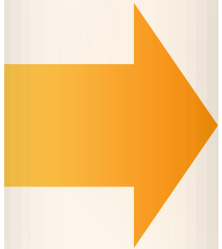
21.80% fewer **classes**

37.81% fewer **properties**

Methodology

Rigorous Annotation Guidelines

```
{
  "sub":"University_Bath",
  "rel":"staff",
  "obj":"2900"
}
```



```
{
  "sub":"University of Bath",
  "subType":"University",
  "rel":"academicStaff",
  "obj": 2900 ,
  "objType":"integer"
}
```

Key improvements

Canonical Names	`University_Bath` → `University of Bath`
Explicit Typing	New `subType` and `objType` fields.
Specific Relations	`staff` (vague) → `academic` (specific)
Correct Data Types	String `"2900"` → integer `2900`

Computing Diff

Knowledge Graph Statistics

Ontology	Text2KGBench-LettrIA		Text2KGBench	
	Sentences	Triples	Sentences	Triples
airport	79 / 227 / 273	260 / 702 / 989	79 / 227	237 / 714
artist	84 / 302 / 198	256 / 896 / 638	84 / 302	252 / 896
astronaut	68 / 86 / 414	266 / 264 / 985	68 / 86	279 / 241
athlete	107 / 186 / 314	304 / 568 / 811	107 / 186	299 / 575
building	103 / 172 / 328	276 / 593 / 956	103 / 172	309 / 588
celestialbody	72 / 122 / 378	203 / 329 / 885	72 / 122	223 / 373
city	217 / 131 / 369	1289 / 479 / 1038	217 / 131	651 / 398
comics character	36 / 66 / 434	92 / 165 / 934	36 / 66	107 / 215
company	56 / 97 / 403	174 / 314 / 928	56 / 97	157 / 300
film	127 / 137 / 363	368 / 369 / 622	127 / 137	378 / 398
food	153 / 245 / 255	473 / 683 / 681	153 / 245	532 / 734
mean of transportation	92 / 222 / 278	271 / 646 / 745	92 / 222	276 / 647
monument	19 / 73 / 427	64 / 343 / 1365	19 / 73	55 / 293
musicalwork	209 / 81 / 419	842 / 285 / 912	209 / 81	604 / 221
politician	135 / 184 / 316	415 / 688 / 1089	135 / 184	424 / 550
scientist	149 / 110 / 390	387 / 259 / 559	149 / 110	411 / 300
sportsteam	110 / 125 / 375	375 / 369 / 1294	110 / 125	401 / 375
university	71 / 85 / 415	337 / 228 / 749	71 / 85	283 / 248
writtenwork	127 / 195 / 305	267 / 628 / 861	127 / 195	381 / 557
TOTAL	**2014 / 2846 / 6654**	**6919 / 8808 / 17101**	**2014 / 2846**	**6259 / 8623**

Number of sentences and triples for each benchmark

Split into test / train / train-ext for T2KGBench-LettrIA

New extended train dataset (synth. generated)

Evaluation

Experimental Setup

Proprietary Models
Zero-Shot

Google
Gemini 2.0, 2.5
Flash Lite, Flash, Pro

OpenAI
GPT-4.1, GPT-4o
Full, Mini, Nano

Anthropic
Claude 3, 3.5, 3.7
Haiku, Sonnet, Opus

Open-Weights Models
Fine-Tuned

Mistral Small 3.2
24B

Gemma 3
4B, 12B, 27B

Phi-4
14B

Qwen 3
from 0.6B to 32B

FINE-TUNING CONFIGURATION

Classic
Trained on the new, curated training set (SFT, LoRA)

Extended
Augmented training set with synthetic data
+500 sentences / ontology

Generalization
Leave-one-ontology-out setting (*city ontology*)

Evaluation

Key Result 1: Fine-Tuning is Paramount

Smaller, fine-tuned open-weights models massively outperform larger, proprietary models.

Category	Model	Entity F1
Fine-Tuned	**Mistral-Small-3.2 (ext.)**	**0.8837**
Fine-Tuned	Gemma-3-27B-it	0.8680
Zero-Shot	gpt-4.1-mini	0.6866
Zero-Shot	**gemini-2.5-pro**	**0.6595**

High-quality, schema-aligned training data is more critical than raw model size.

Evaluation

Key Result 2: Generalization, Reliability and Efficiency

- **Fine-tuning** teaches true generalization, not just memorization.
  - Fine-tuned models maintain their performance advantage even on an unseen ontology.
  - gemma-3-12b-it (F1 **0.8376**) significantly outperforms the best proprietary model, claude-sonnet-4 (F1 **0.7829**).
- **Fine-tuned models** achieve near-perfect (> 99%) structural validity and have extremely low hallucination rates.
- **Proprietary models** show variable reliability; some are strong (gemini-2.5-pro), while others fail catastrophically.
- **Efficiency** (in terms of latency) varies by orders of magnitude
  [2s (gemini-flash) - 27s (claude-opus)] versus <0.02s (fine-tuned) models
  which is directly correlated with **cost**

Lesson Learned

A Fundamental Trade-Off

- **Specialization Wins**
  For complex, structured tasks like Text2KG, task-specific fine-tuning is the most effective strategy.

- **Fine-Tuning Teaches Generalization**
  The process instills a deep, transferable understanding of the task's logic.

- **Accessibility vs. Efficiency**
  - **Proprietary APIs**
    Invaluable for rapid prototyping. Higher latency and pay-per-call costs.
  - **Fine-Tuned Models**
    A strategic investment. Delivers superior performance, speed, and economics for high-volume applications.

lettria

Fall Launch 2025

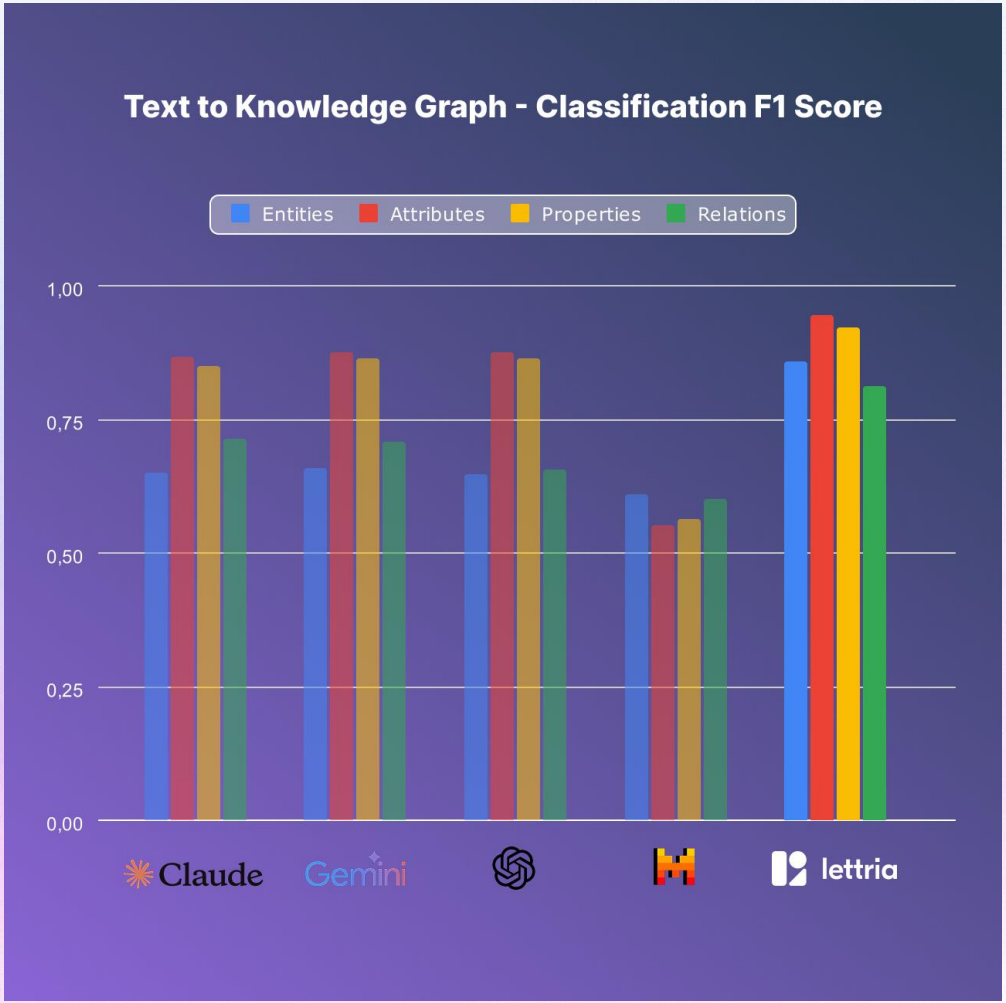
Empowering Graph-Based Agents with our new model Lettria Perseus

We benchmarked Lettria Perseus against leading LLMs on **attribute extraction tasks** central to building **reliable graph-based agents**. The results confirm that our specialized model **delivers higher precision** and recall than general-purpose alternatives, ensuring more **accurate and trustworthy document intelligence for enterprise use**.

Meet Perseus

https://www.lettria.com/benchmarks/benchmark-empowering-graph-based-agents-with-lettrias-new-model

https://deepmind.google/models/gemma/gemmaverse/lettria/



Text-to-Graph Leaderboard

Compare reliability, extraction quality, hallucination rates, latency, and cost across all top proprietary APIs and fine-tuned open-weights models. Explore the full benchmark results and identify the best model profile for your workload.

Start building Scientific paper

https://perseus.lettria.com/leaderboard

Best reliability
100.00%

Fastest average latency
1.36s

Lowest average cost
$0.01

Search model name...

Sort by average f1

Rank	Model	Output Reliability (%)	Average F1 ↓	Entities F1 (Recognition)	Entities F1 (Typing)	Entities F1 (Extract
1	Fine-tuned Mistral Small 3.2	99.95	0.89	0.87	0.90	0.92
2	Fine-tuned Gemma 3 27B Inst	99.8	0.89	0.86	0.89	0.92
3	Fine-tuned Qwen 3 32B	99.9	0.88	0.86	0.90	0.92

Free tier access to PERSEUS

- → https://github.com/Lettria/perseus-client
- → https://docs.perseus.lettria.net/
- → https://app.perseus.lettria.net/app/graphs

Text to Graph:
From General LLMs
to Lettria Perseus

ECLADATTA

ExtraCtion of LAtent knowledge in Documents by conjointly Analyzing Texts and TAbles

Orange, EURECOM, IRIT

Github

TRIPLET Workshop @ ESWC 2026

TRIPLET 2026: International Workshop on Extraction from Triplet Text-Table-Knowledge Graph and associated Challenge @ 23rd European Semantic Web Conference (ESWC 2026), Dubrovnik, Croatia

Room 2

Programme - Monday 11 May 2026

Time	Session	Title	Authors / Details
09:00	Opening	Welcome & Introduction	Organisers
09:10	Keynote	**Leave the modalities alone? When, where and if we should extract knowledge graphs from multiple modalities**	Paul Groth
10:10	Paper Session	**Towards Foundation Models for Relational Databases with Language Models and Graph Neural Networks**	Jingcheng Wu, Ratan Bahadur Thapa, Mojtaba Nayyeri, Lucas Etteldorf, Max Finkenbeiner, Fabian Leeske and Steffen Staab
10:30	Break	Coffee Break	
11:00	Paper Session	**From Rows to Narratives: Benchmarking Semantic Relatedness Across Tables and Paragraphs**	Fanfu Wei, Thibault Ehrhart and Raphael Troncy
11:20	Challenge Session	**TRIPLET Challenge Overview**	Raphael Troncy, Yoan Chabot, Véronique Moriceau, Nathalie Aussenac-Gilles and Mouna Kamel
11:30	Challenge Session	**Task 1: Iterating on text-table pairing under prosumer hardware constraint**	Clement Benesse
11:40	Challenge Session	**Task 1: Candidate-Aware Table Serialization and Cross-Encoder Ranking for Table–Text Relatedness**	Yuuki Tachioka
11:50	Challenge	**Task 2: TripleMap: Hybrid Schema-Grounded**	Shanthini Malarvizhi

Take Away

- Text2KG is moving from extracting triples to constructing situated, evidence-grounded, schema (ontology)-aware, multimodal, and evaluable knowledge graph
- LLMs have made Text2KG easier to demo but harder to trust
- What does it mean for a generated KG to be correct?

LettRAGraph

Let’s discuss!

Credits: Pasquale Lisena, Thibault Ehrhart, Ismail Harrando, Julien Plu, Thomas Schleider, Youssra Rebboud