KE 4 TXT || NLP 4 KGC

Knowledge Extraction (KE)

TASK: TO POPULATE A KNOWLEDGE BASE WITH INFORMATION FROM EXISTING LEGACY RESOURCES

WHEN KE FROM TEXT, WE USE NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES

Natural Language Processing (NLP)

- Branch of Computer Science at the intersection with Artificial Intelligence and (Computational) Linguistics
- Linguistics: the scientific study of language and its structure, including the study of grammar, syntax, and phonetics …
- Computational Linguistics: studies language through knowledge representation methods. Develops models of language …

NLP in a nutshell

More pointers than anything else...

NLP Applications

Search

Translation

Generation

Summarisation

Chatbots

Knowledge extraction

NLP Tasks

- Tokenisation
- Lemmatisation
- Part-of-speech (POS) tagging
- Stemming
- Sentence parsing / grammar induction
- Co-reference resolution
- Word sense disambiguation (WSD)
- Named Entity Recognition (NER)
- Named Entity Classification
- Entity Linking (EL)
- Sentiment Analysis
- Relation Extraction
- Semantic role labelling (SRL)
- […]

NLP (meta) tasks

Generation

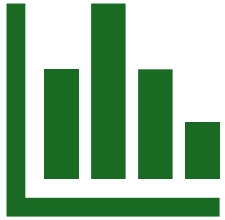
Classification

Annotation

NLP Research

Symbolic

(~1950~1990)



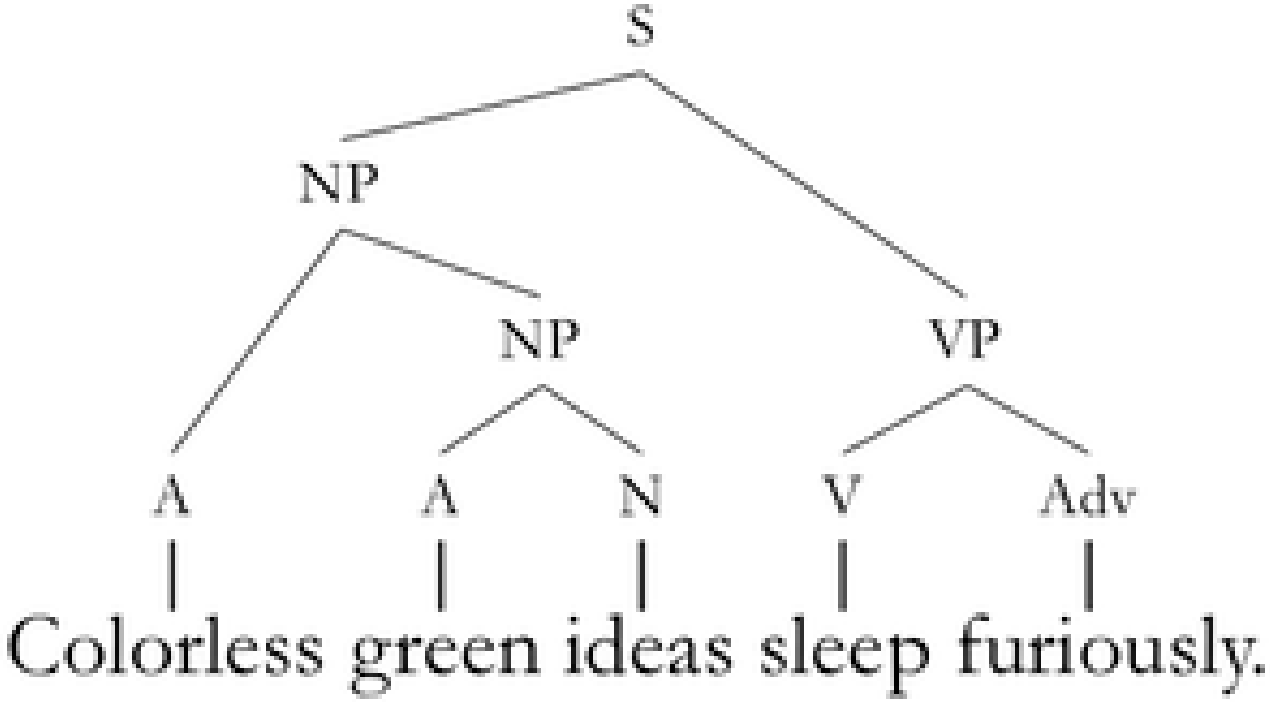
Statistical

(~1990~2010)

Neural

(~2010~today)

Symbolic approaches
~
formal grammars
~
rules
~
reference data



PRINCETON UNIVERSITY

WordNet

A Lexical Database for English

Term Frequency - Inverse Document Frequency (TF-IDF)

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

TF: Il peso di un termine che ricorre in un documento è semplicemente proporzionale alla frequenza del termine. Hans Peter Luhn (1957)

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

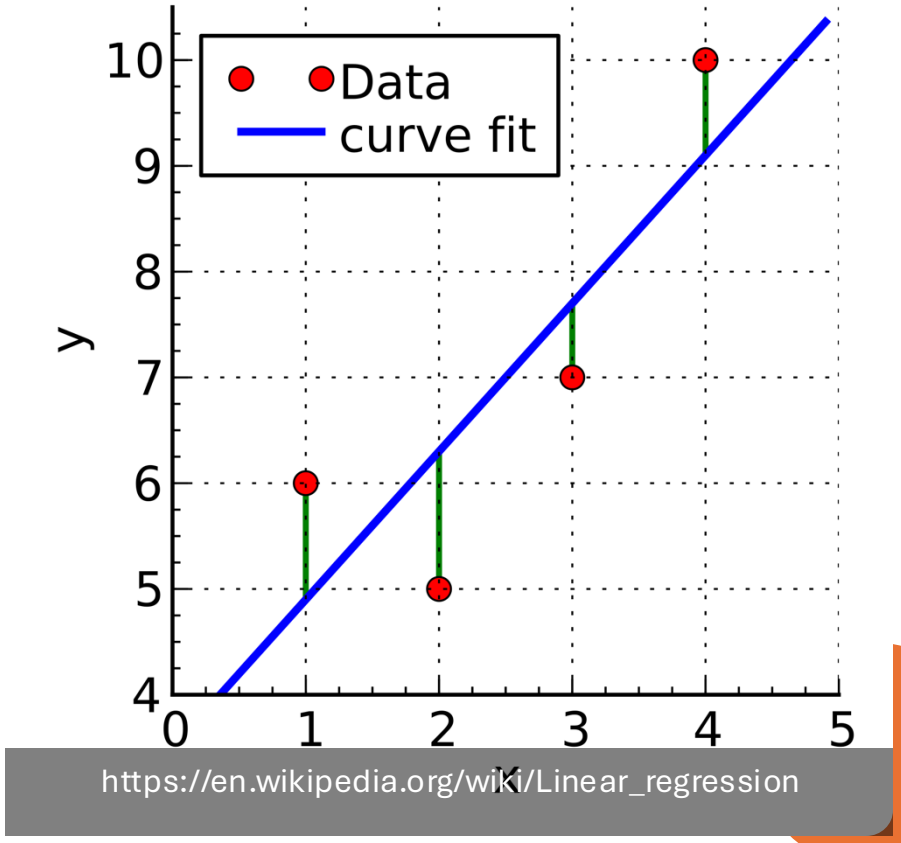
La specificità di un termine può essere quantificata come una funzione inversa del numero di documenti in cui compare. Karen Spärck Jones (1972)

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d : d \in D \text{ and } t \in d\}|}$$

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

Machine learning

- Support Vector Machine (SVM)
- Linear regression
- Logistic regression
- Decision tree (…symbolic?)
- …



https://en.wikipedia.org/wiki/Linear_regression

Abstract Meaning Representation (AMR)

Example [edit]

Example sentence: *The boy wants to go.*

```
(w / want-01
  :arg0 (b / boy)
  :arg1 (g / go-01
    :arg0 b))
```

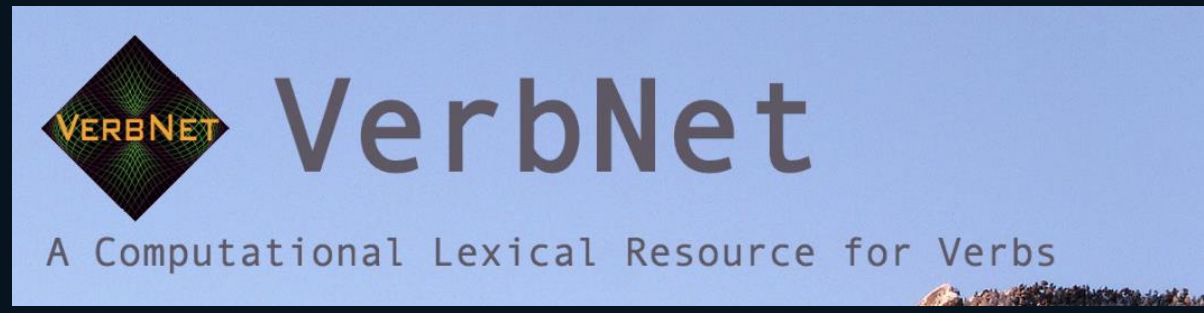
Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. "Abstract meaning representation for sembanking." In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pp. 178-186. 2013.

https://en.wikipedia.org/wiki/Abstract_Meaning_Representation

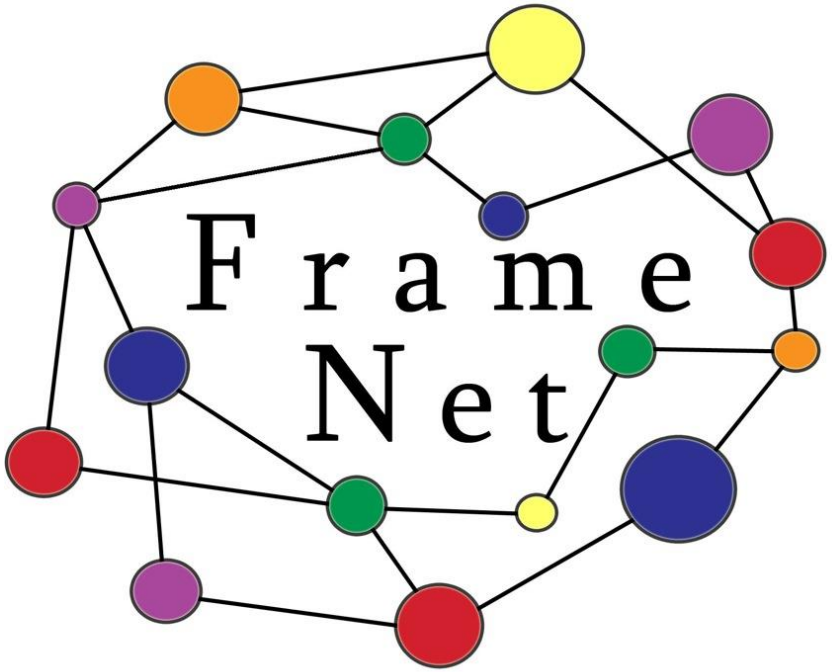
Word2Vec

“Word2vec is a technique in natural language processing (NLP) for obtaining vector representations of words. These vectors capture information about the meaning of the word based on the surrounding words. The word2vec algorithm estimates these representations by modeling text in a large corpus. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. Word2vec was developed by Tomáš Mikolov and colleagues at Google and published in 2013.“ (Wikipedia)

NLP Resources



FrameNet



https://framenet.icsi.berkeley.edu/

GLOBAL FRAMENET

The hub for collaboration in the development of frame-based language resources and applications

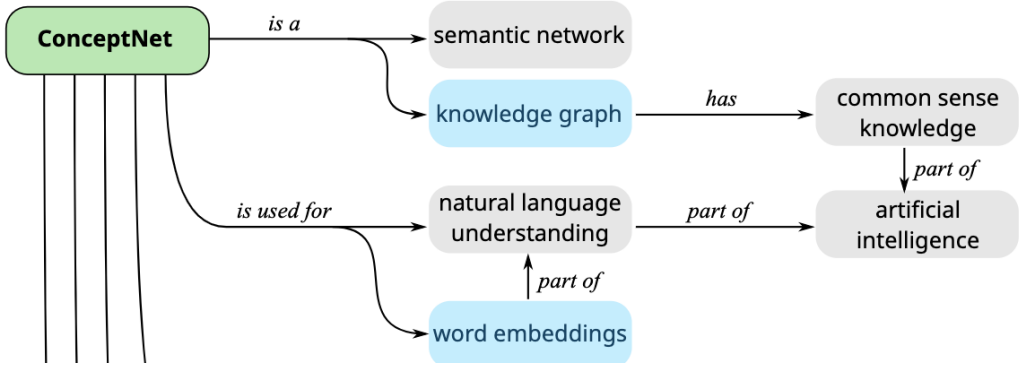
https://www.globalframenet.org/

FrameNet maps meaning to form in contemporary English through the theory of Frame Semantics.

ConceptNET

ConceptNet

An open, multilingual knowledge graph



ConceptNet *is a freely-available semantic network, designed to help computers understand the meanings of words that people use.*

https://conceptnet.io/

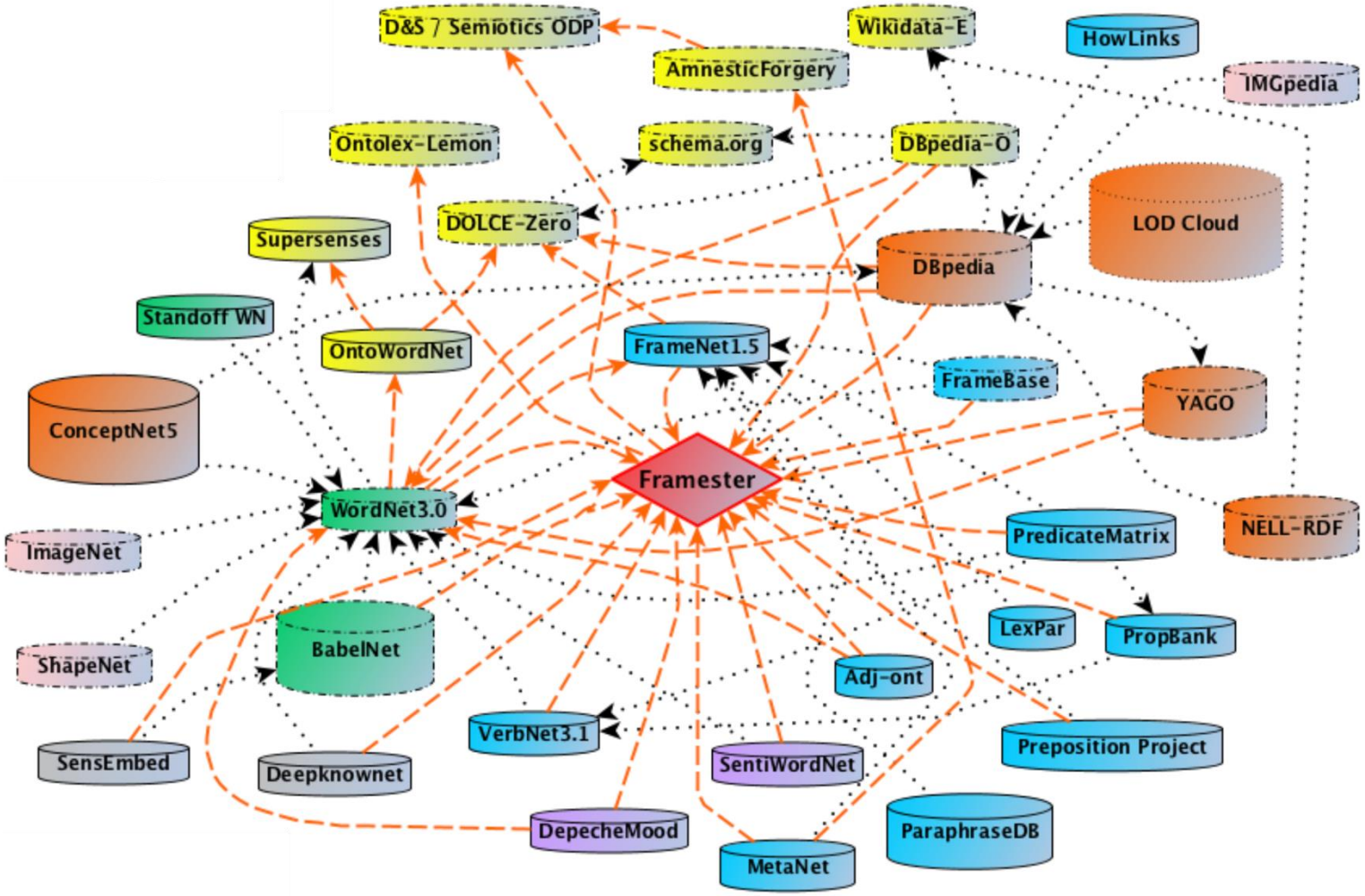
DBpedia Spotlight

Shedding light on the web of documents

It is a tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia.

http://demo.dbpedia-spotlight.org/

FRAMESTER



Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti and Diego Reforgiato Recupero. Framester: A Wide Coverage Linguistic Linked Data Hub. In: *Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016). Bologna, Italy, 2016* DOI: 10.1007/978-3-319-49004-5_16

```
PREFIX wn30instances:
PREFIX wn30schema:
PREFIX depmood:
SELECT * WHERE {
        ?syn depmood:AFRAIDscore ?afraid
        depmood:AMUSEDscore ?amused ;
        depmood:ANGRYscore ?angry ;
        depmood:ANNOYEDscore ?annoyed ;
        depmood:DONT_CAREscore ?dontCare
        depmood:HAPPYscore ?happyScore ;
        depmood:INSPIREDscore ?inspired ;
        depmood:SADscore ?sad
}
LIMIT 10
```

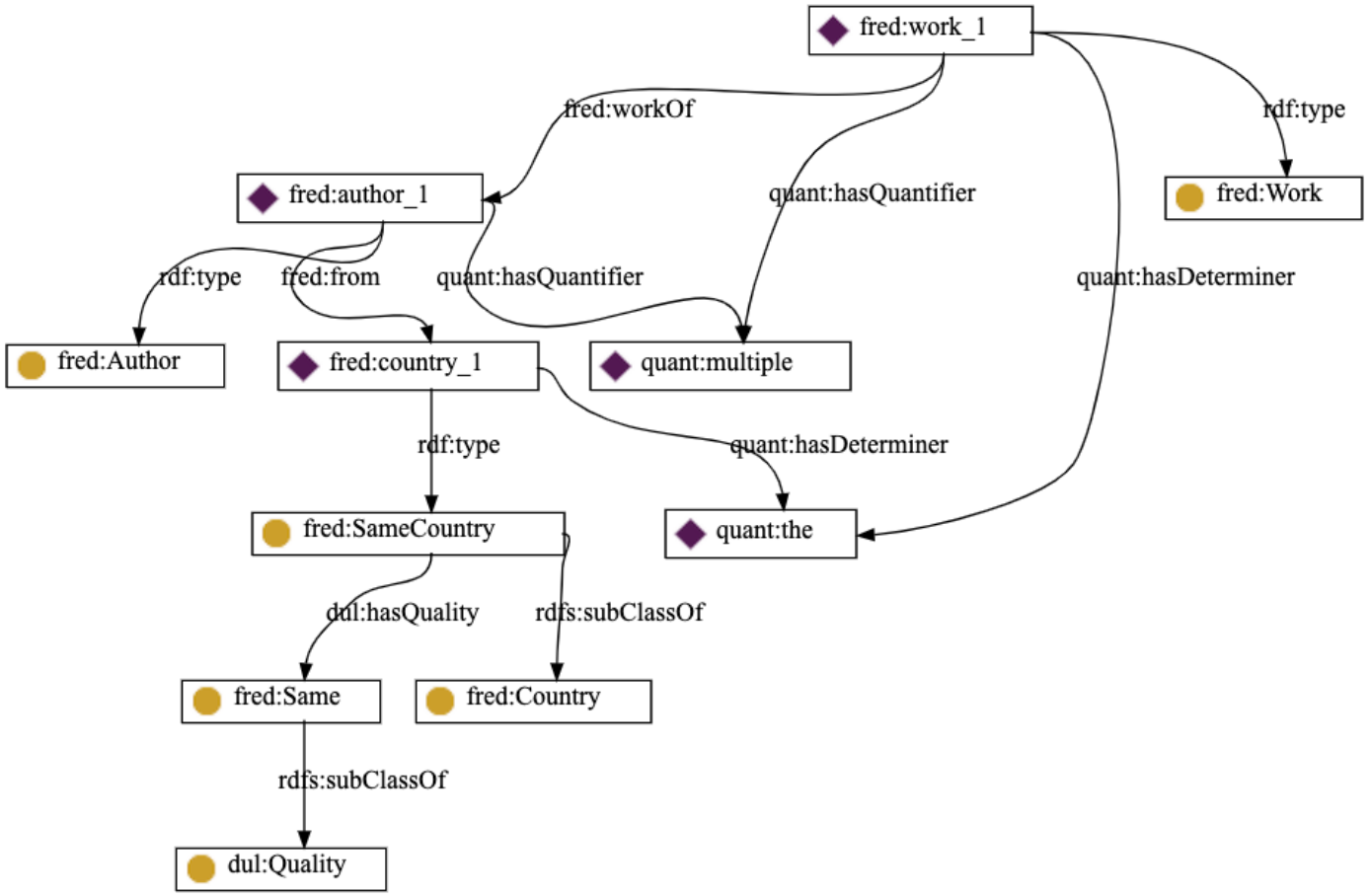
FRED

Machine Reading for the Semantic Web

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

Gangemi, Aldo, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. "Semantic web machine reading with FRED." *Semantic Web* 8, no. 6 (2017): 873-893.

- Which are the works of authors from the same country?



Bevilacqua, Michele, Rexhina Blloshmi, and Roberto Navigli. "One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, pp. 12564-12573. 2021.

- https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Introduces the “transformer architecture”, at the basis of most Large Language Models

Popular large language models

From sources across the web

- Llama
- BERT
- Claude
- Sentiment analysis
- Ernie
- Summarization
- Google
- Mistral
- Cohere
- Gemini
- LaMDA
- Translation
- Language modeling
- Vicuña
- Falcon
- StableLM
- GPT-4
- GPT-3
- OpenAI
- BLOOM
- Orca
- Databricks and Mosaic
- PaLM
- Content generation

Problems with LLMs

- Bias
- Hallucinations
- Outdated knowledge
- Lack of transparency
- Lack of accountability
- Lack of consistency
- Intellectual property
- Privacy

Bates, Jo, Elli Gerakopoulou, and Alessandro Checco. "Addressing labour exploitation in the data science pipeline: views of precarious US-based crowdworkers on adversarial and co-operative interventions." *Journal of Information, Communication and Ethics in Society* 21, no. 3 (2023): 342-357.

Goetze, Trystan S. "AI Art is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks." In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 186-196. 2024.

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. "Accountability in artificial intelligence: what it is and how it works." *Ai & Society* 39, no. 4 (2024): 1871-1882.

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023

Last updated: 18-06-2024 - 16:29

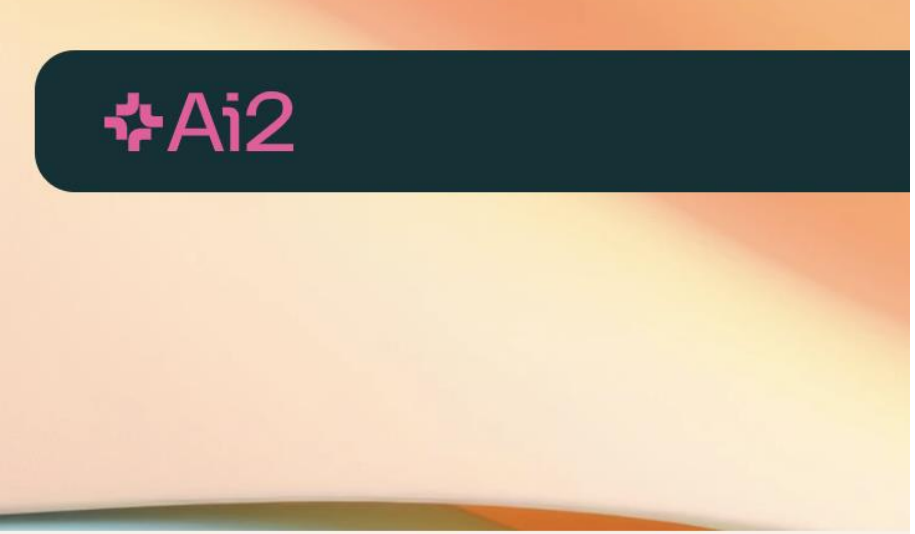
6 min read

https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Some ways to use LLMs

- Fine-tuning
- Zero-shot
- In-context learning
  - One-Shot Learning
  - Few-Shot Learning
- Chain of thought

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.



https://allenai.org/olmo

OLMo 2 is a family of fully-open language models, developed start-to-finish with open and accessible training data, open-source training code, reproducible training recipes, transparent evaluations, intermediate checkpoints, and more.

Groeneveld, Dirk, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha et al. "Olmo: Accelerating the science of language models." *arXiv preprint arXiv:2402.00838* (2024).

Let’s try it out!

You are an ontology engineer and need to design an ontology. You use Competency Questions to support the ontology design activity. List for me the classes and properties that could answer the following Competency Question: "Which are the works of authors from the same country?"

You are an ontology engineer and need to design an ontology. You use Competency Questions to support the ontology design activity. List for me the classes and properties that could answer the following Competency Question: "Which are the works of authors from the same country?". ***Reply with only the list of classes and properties, in a JSON structure.***

Python libraries

Spacey:
https://spacy.io/

NLTK:
https://www.nltk.org/

Huggingface:
https://huggingface.co/

SKLearn:
https://scikit-learn.org/stable/

Links

- Verbnet: https://verbs.colorado.edu/verbnet/
- Wordnet: https://wordnet.princeton.edu/
- Propbank: https://propbank.github.io/
- FrameNet: https://framenet.icsi.berkeley.edu/ https://www.globalframenet.org/
- NLTK + Propbank example: https://www.nltk.org/howto/propbank.html
- AMRLIB: https://spacy.io/universe/project/amrlib
- Framester: http://etna.istc.cnr.it/framester_web/
- FRED: http://wit.istc.cnr.it/stlab-tools/fred/demo/?

Bibliography

- Orlando, Riccardo, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. "ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget." arXiv preprint arXiv:2408.00103 (2024).
- Bevilacqua, Michele, Rexhina Blloshmi, and Roberto Navigli. "One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 14, pp. 12564-12573. 2021.
- Speer, Robyn, Joshua Chin, and Catherine Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge." In Proceedings of the AAAI conference on artificial intelligence, vol. 31, no. 1. 2017.
- Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti and Diego Reforgiato Recupero. Framester: A Wide Coverage Linguistic Linked Data Hub. In: Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016). Bologna, Italy, 2016 DOI: 10.1007/978-3-319-49004-5_16
- Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. "Abstract meaning representation for sembanking." In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, pp. 178-186. 2013.
- Karen Sparck Jones, A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL, in Journal of Documentation, vol. 28, n. 1, 1972-01, pp. 11–21, DOI:10.1108/eb026526. URL consultato l'8 luglio 2023.
- H. P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, in IBM Journal of Research and Development, vol. 1, n. 4, 1957-10, pp. 309–317, DOI:10.1147/rd.14.0309. URL consultato l'8 luglio 2023.
- Manning, Christopher, and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.
- Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).