Natural Language Processing - 2nd Semester (2024-2025)
1038141

1.2 - Introduction to NLP

SAPIENZA
UNIVERSITÀ DI ROMA

Prof. Stefano Faralli
faralli@di.uniroma1.it

Prof. Iacopo Masi
masi@di.uniroma1.it

**credits are reported in the last slide

Before we start

Before we start

SAPIENZA
UNIVERSITÀ DI ROMA

1.2 Introduction to NLP

- What is Natural Language Processing?
- Milestones in NLP
- A small bite of Text Representation
- The #BenderRule
- NLP Tasks and Research
- Resources and References
- Q&A

What is Natural Language Processing (NLP)?

Natural Language Processing (NLP):

is an **interdisciplinary field** concerned with the **interactions** between **computers** and **natural human languages** (e.g. English, Italian, Spanish, Chinese etc..) — **speech or text.**

What is Natural Language Processing (NLP)?

NLP-powered software helps us in our daily lives in various ways, for example:

- **Personal voice assistants:** e.g., Siri, Cortana, and Google Assistant.
- **Auto-complete:** In search engines, e.g., Google.
- **Spell checking:** Almost everywhere, in your browser, your IDE (e.g. Visual Studio), desktop apps (e.g. Microsoft Word), Grammarly ….
- **Machine Translation:** Google Translate, …

What is Natural Language Processing (NLP)?

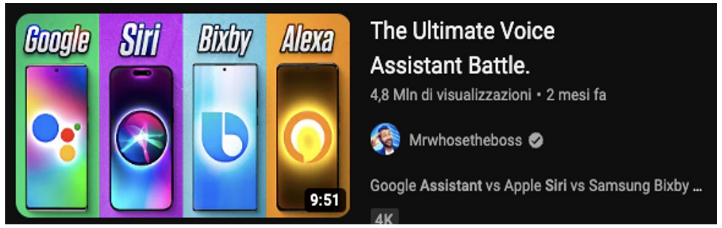
NLP-powered software helps us in our daily lives in various ways, for example:

- **Personal voice assistants:** e.g., Siri, Cortana, and Google Assistant.
- **Auto-complete:** In search engines, e.g., Google.
- **Spell checking:** Almost everywhere, in your browser, your IDE (e.g. Visual Studio), desktop apps (e.g. Microsoft Word), Grammarly ….
- **Machine Translation:** Goog

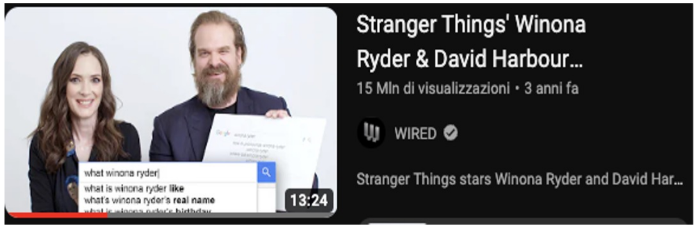
….

Can you provide more examples of NLP-powered software and applications?

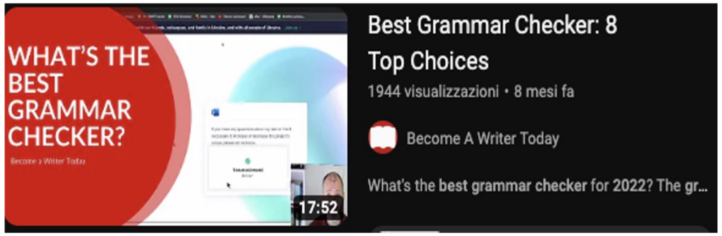
What is Natural Language Processing (NLP)?



https://youtu.be/rqTGKcoq2Os



https://youtu.be/nsEO8LALcyo

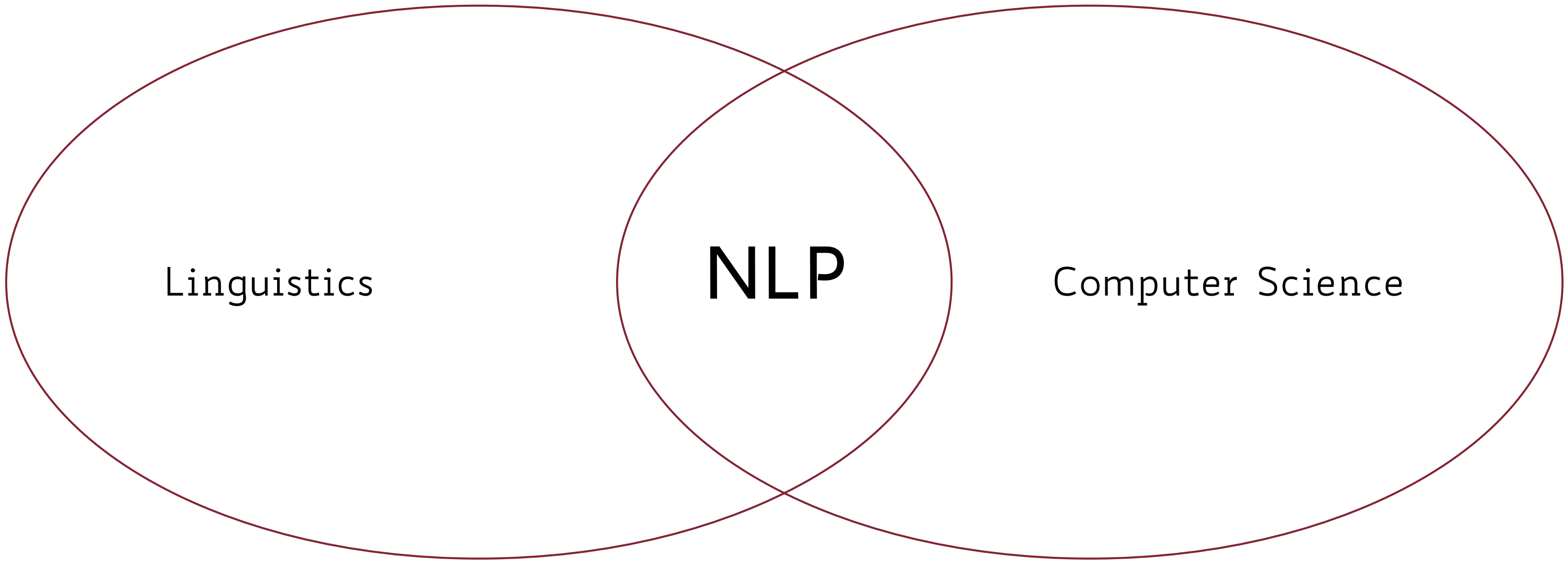


https://youtu.be/qOz-6Ahgf8U

https://youtu.be/jxP_k9jE_sU

What is Natural Language Processing (NLP)?

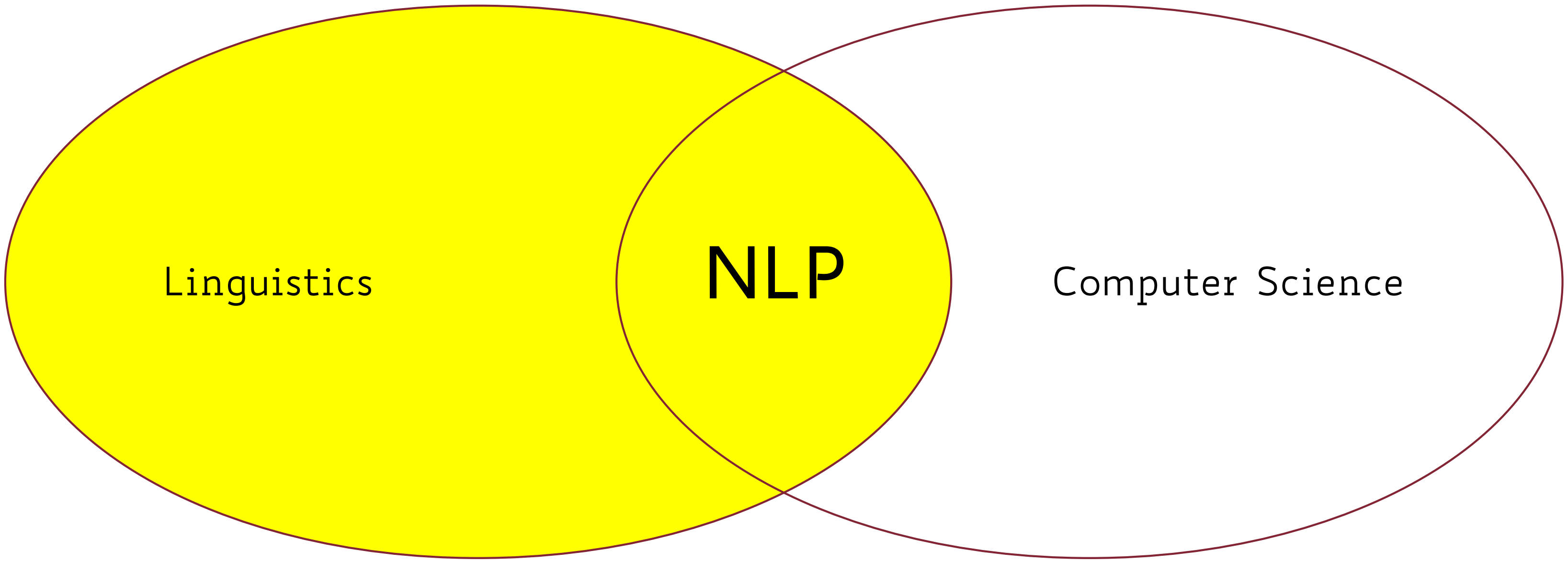
NLP is mainly divided into two fields: **Linguistics** and **Computer Science**.



[TowardsDataScience1]

What is Natural Language Processing (NLP)?

The **Linguistics** side focuses on understanding the structure of language,



[TowardsDataScience1]
[Bender, 2013]

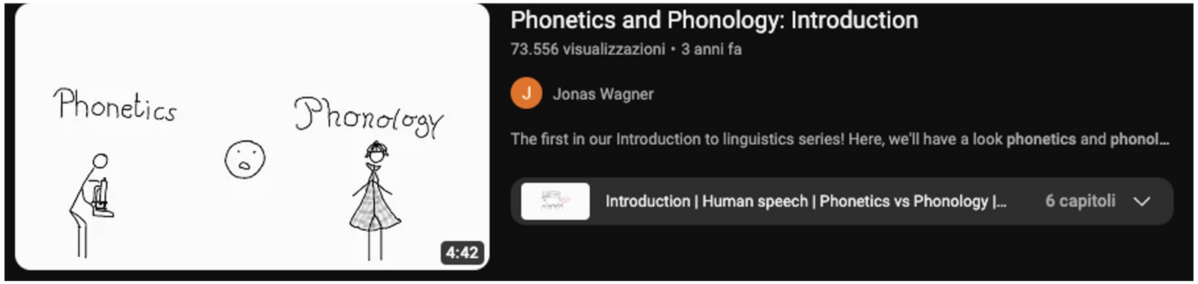
What is Natural Language Processing (NLP)?

The **Linguistics** side focuses on understanding the structure of language, including the following sub-fields [Bender, 2013]:

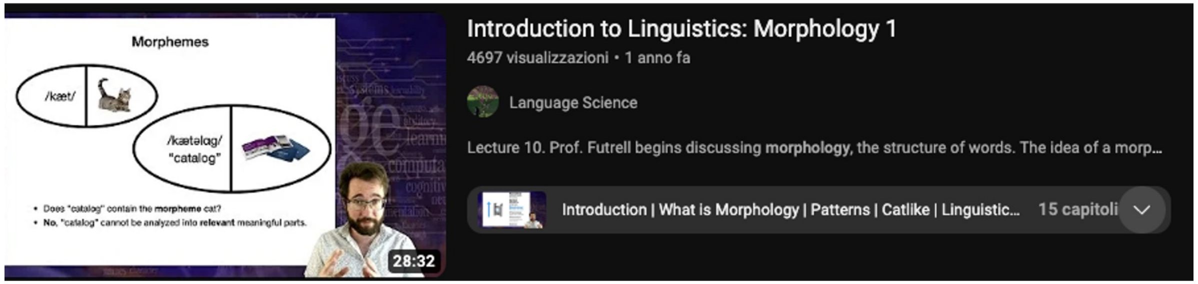
1. Phonetics: The study of the **sounds** of human language.
2. Phonology: The study of the **sound systems** in human languages.
3. Morphology: The study of the **formation** and **internal structure** of **words**.
4. Syntax: The study of the **formation** and **internal structure** of **sentences**.
5. Semantics: The study of the **meaning** of **sentences**.
6. Pragmatics: The study of the way **sentences** with their **semantic** meanings are **used** for particular **communicative goals**.

[TowardsDataScience1]
[Bender, 2013]

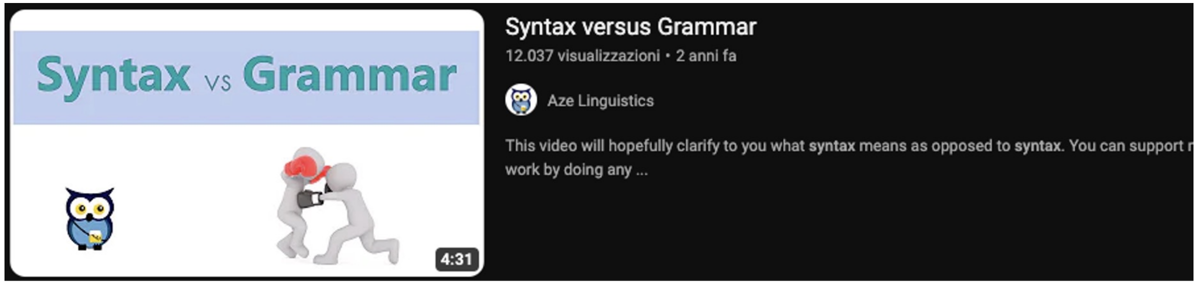
What is Natural Language Processing (NLP)?



https://youtu.be/8Od2CEeMyQQ

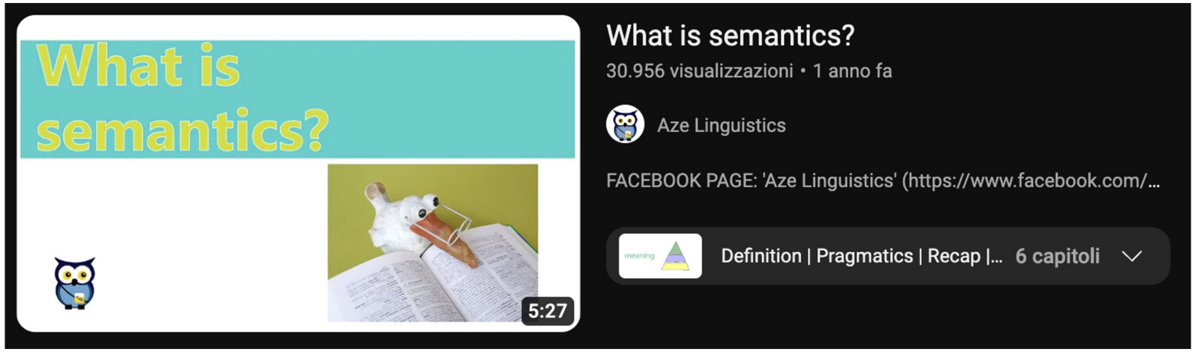


https://youtu.be/MAwSrc6qMTQ

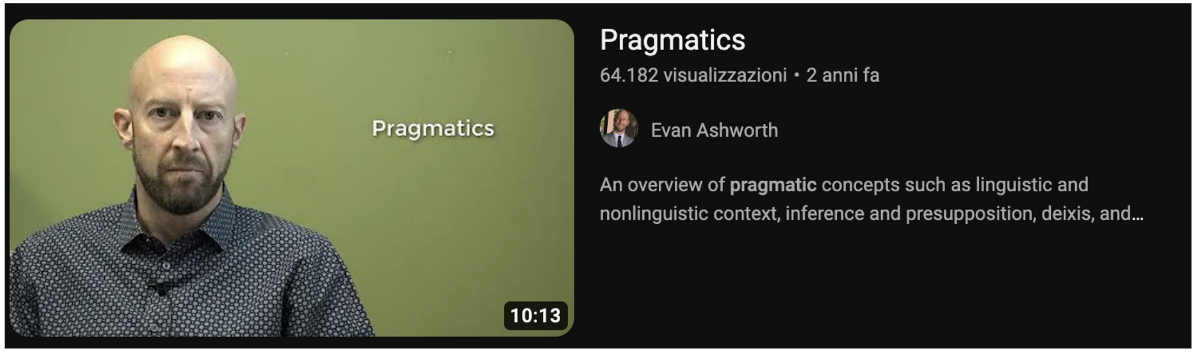


https://youtu.be/o6N_DiTSAA4

What is Natural Language Processing (NLP)?



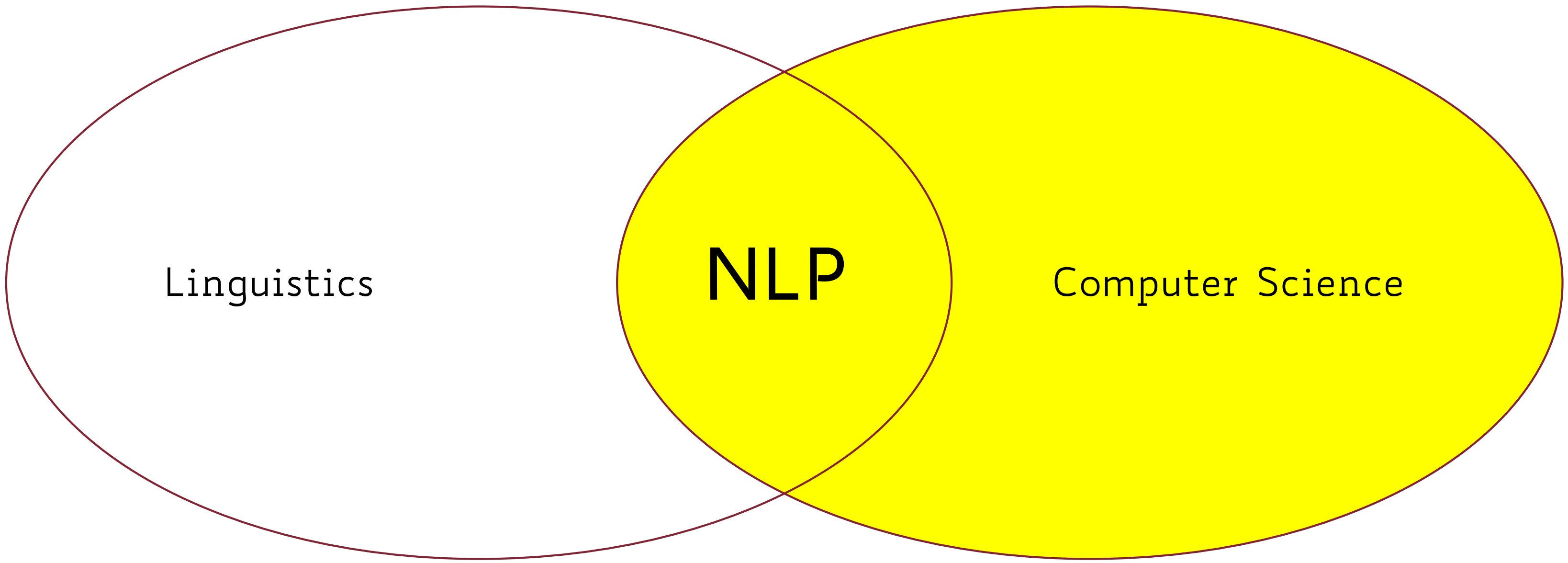
https://youtu.be/SyFoFsuUNZk



https://youtu.be/dsPswzOBsKO

What is Natural Language Processing (NLP)?

The **Computer Science** side is concerned with translating linguistic knowledge and domain expertise into computer programs with the help of sub-fields such as Artificial Intelligence.



[TowardsDataScience1]

What is Natural Language Processing (NLP)?

The **Computer Science** side is concerned with translating linguistic knowledge and domain expertise into computer programs with the help of sub-fields such as Artificial Intelligence.

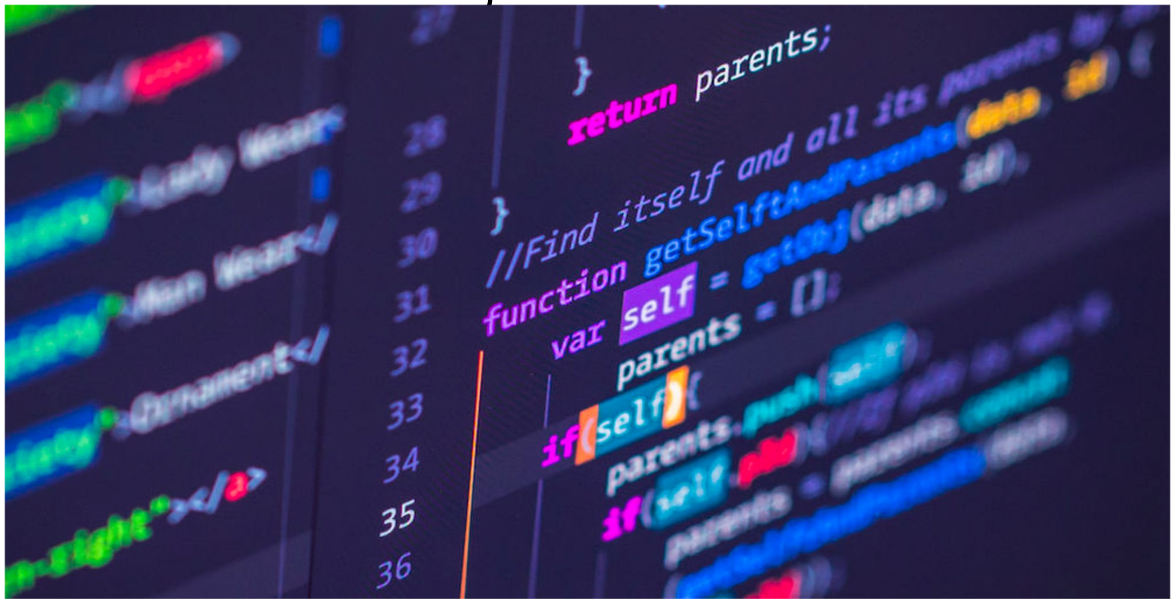
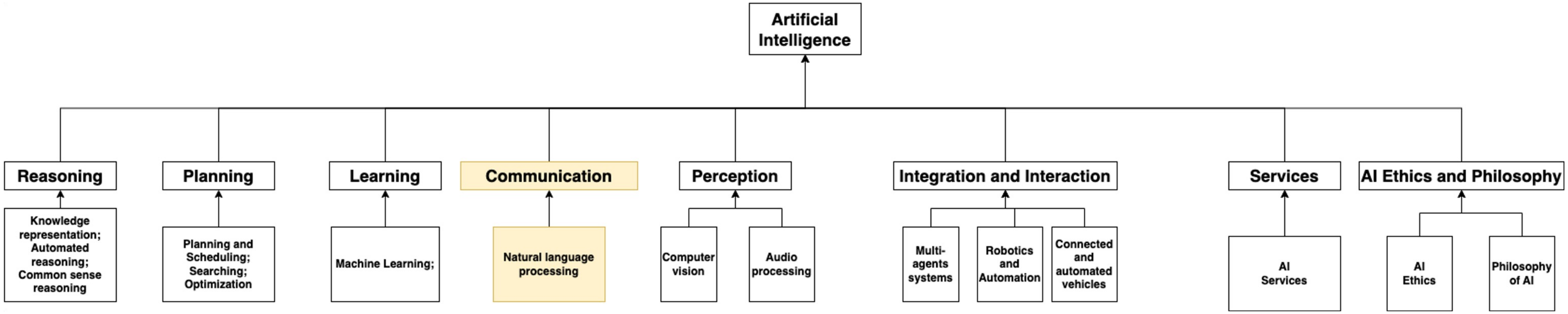


Image Source: https://al.nd.edu/

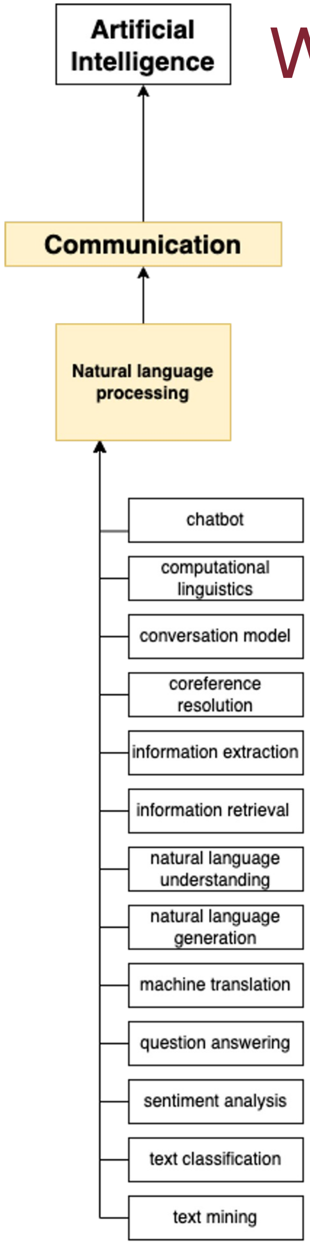
[TowardsDataScience1]

What is Natural Language Processing (NLP)?



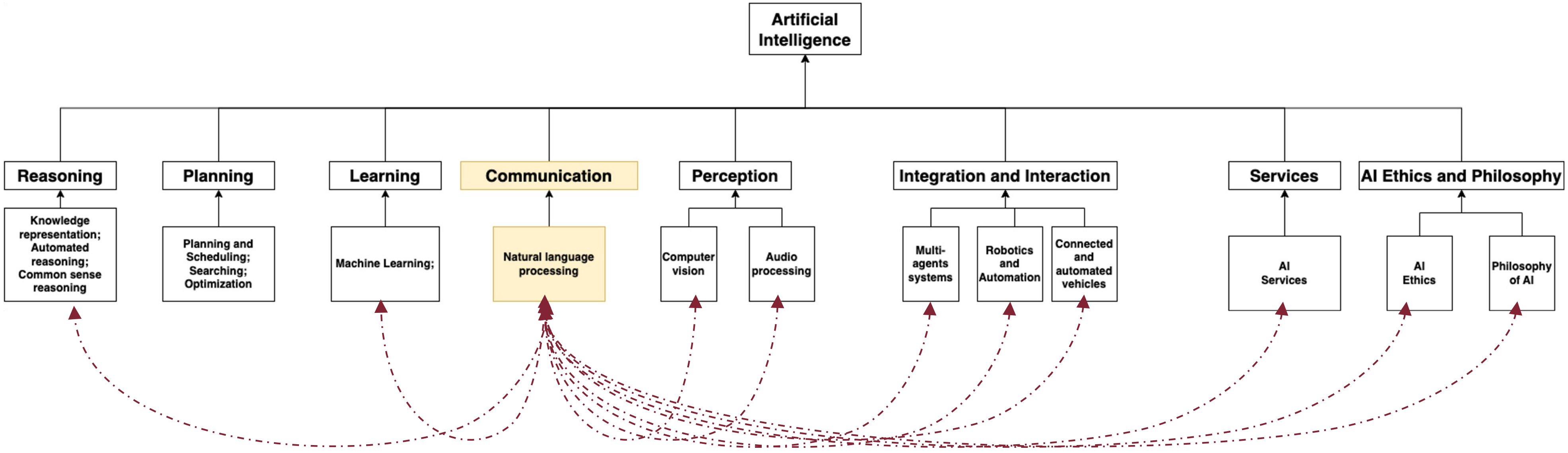
[JCR, 2020] Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B., AI WATCH. Defining Artificial Intelligence, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163. https://publications.jrc.ec.europa.eu/repository/handle/JRC118163

What is Natural Language Processing (NLP)?



… and many more to come!

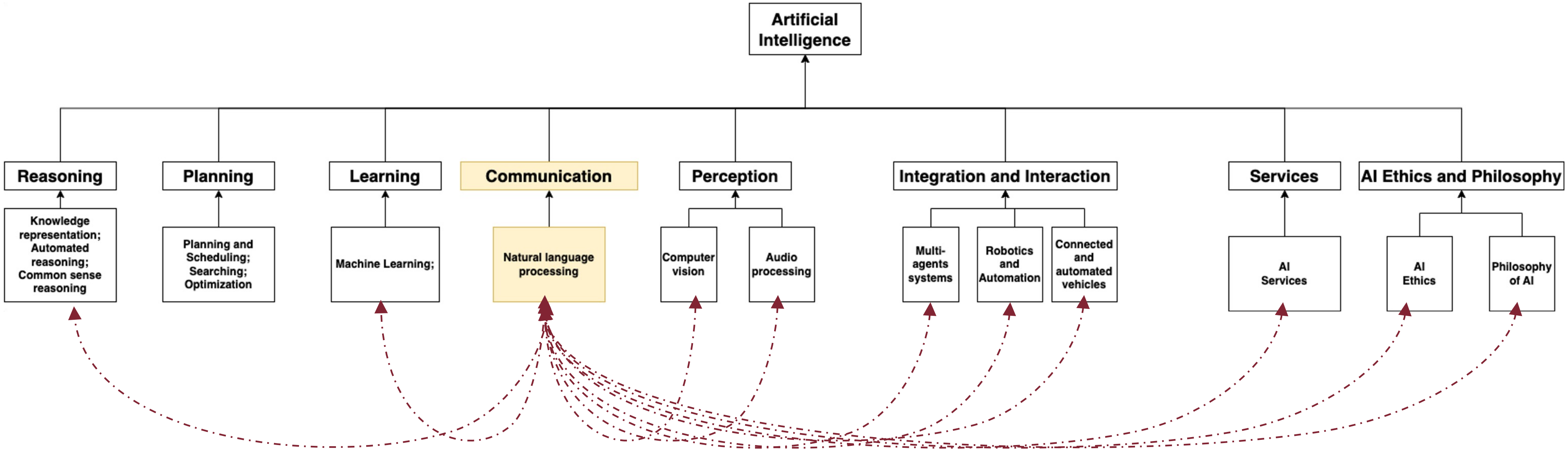
What is Natural Language Processing (NLP)?



"interactions"!!

[JCR, 2020] Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B., AI WATCH. Defining Artificial Intelligence, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163. https://publications.jrc.ec.europa.eu/repository/handle/JRC118163

What is Natural Language Processing (NLP)?



"interactions"!!

Can you provide examples of such "interactions"?

Milestones in NLP

rule-based systems

[TowardsDataScience1]

Milestones in NLP

rule-based systems

not a tombstone!

[TowardsDataScience1]

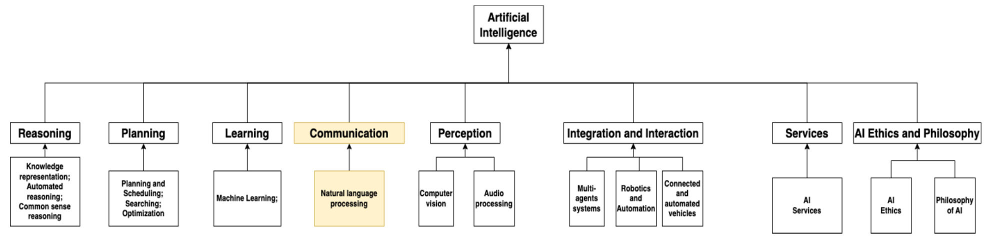
Milestones in NLP

rule-based systems

statistical and classical machine learning models

[TowardsDataScience1]

Milestones in NLP



rule-based systems

statistical and classical machine learning models

deep learning models

[TowardsDataScience1]

Milestones in NLP: rule-based systems

Rule-based systems:

rely heavily on crafting domain-specific rules (e.g., **regular expressions**).

It can automate simple tasks such as extracting structured data (e.g. dates, names) from unstructured data (e.g. webpages, emails).

However, due to the complexity of human languages, rule-based systems aren't robust, hard to maintain, and can't generalize across different domains.

[TowardsDataScience1]

Milestones in NLP: rule-based systems

Rule-based systems:

rely heavily on crafting domain-specific rules (e.g., **regular expressions**).

It can automate simple tasks such as extracting structured data (e.g. dates, names) from unstructured data (e.g. webpages, emails).

However, due to the compl guages, rule-based systems aren't ro n, and can't generalize across different domains.

Can you provide more examples of rule-based systems?

Milestones in NLP: statistical and classical machine learning models approaches

statistical and classical machine learning models approaches:

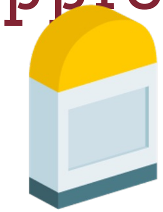
can solve more challenging problems (e.g. spam detection).

Using statistical and probabilistic models to model languages.

Using feature engineering (e.g. **bag of words**, **part of speech tags**) to build machine learning models (e.g. **Naive Bayes**). Those models exploit systematic patterns in training data and can make predictions for unseen data.

[TowardsDataScience1]

Milestones in NLP: classical machine learning models approaches



Classical machine learning models approaches:

can solve more challenging problems (e.g. spam detection).
Using feature engineering (e.g. **bag of words**, **part of speech tags**) to build machine learning models (e.g. **Naive Bayes**).

Those models exploit systematic patterns in training data and can make predicti

Can you provide more examples of features? What do "Naive Bayes models" means?

Milestones in NLP: deep learning models approaches

Deep learning models approaches:

are currently the most popular in NLP research and applications.
They **generalize** even better than the classical machine learning approaches. It doesn't need hand-crafted features or feature engineering because they automatically work as feature extractors, enabling end-to-end model training.
Deep learning models learning capabilities are more powerful than shallow/classical ML ones, which paved its way to achieving the highest scores on various challenging NLP tasks (e.g. Machine Translation).

[TowardsDataScience1]

Deep learning models approaches:

are currently the most popular in NLP research and applications.

They **generalize** even better than the classical machine learning approaches. It doesn't need hand-crafted features or feature engineering because they automatically work as feature extractors, enabling end-to-end model training.

Deep learning models lear … more powerful than shallow/class … paved its way to achieving th … ous challenging NLP tasks (e.g

What do "deep learning" and "generalize" mean?

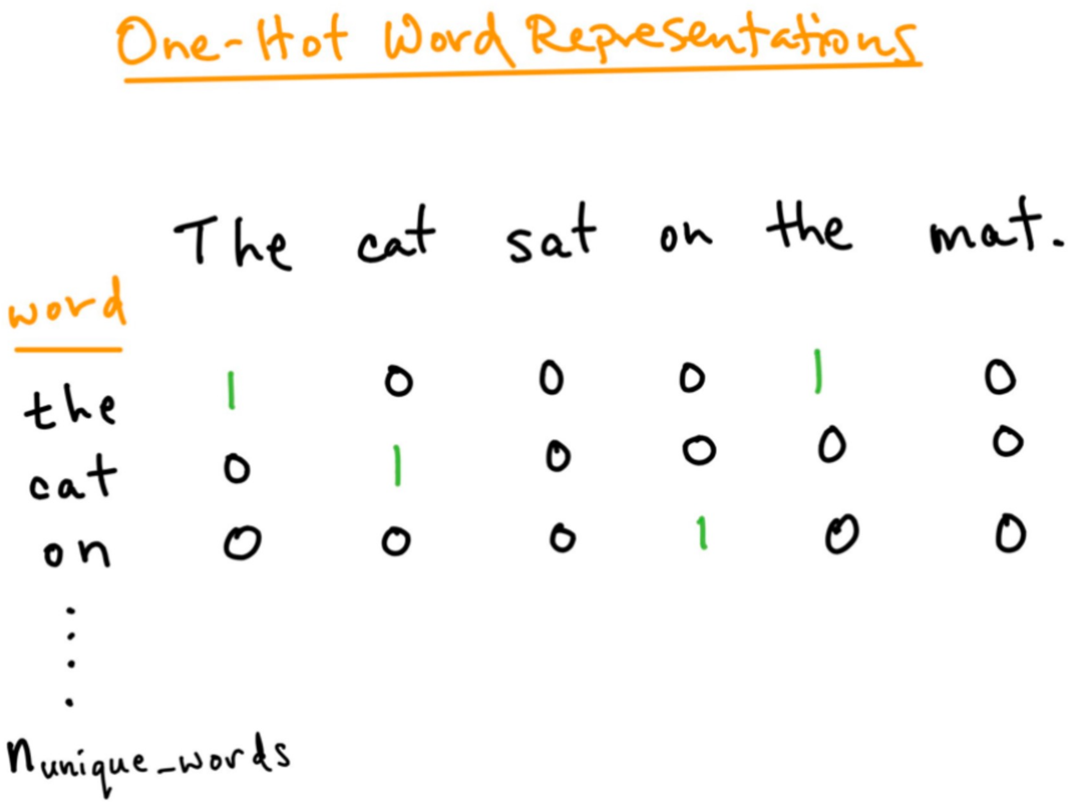
A small bite of Text Representations

In the classical NLP/ML era (before deep learning), text representation techniques were mainly built on a basic idea: **one-hot encodings**, where a sentence is represented by a **matrix** of shape (N x N), where N is the number of unique tokens in the sentence.

[TowardsDataScience1]

A small bite of Text Representations

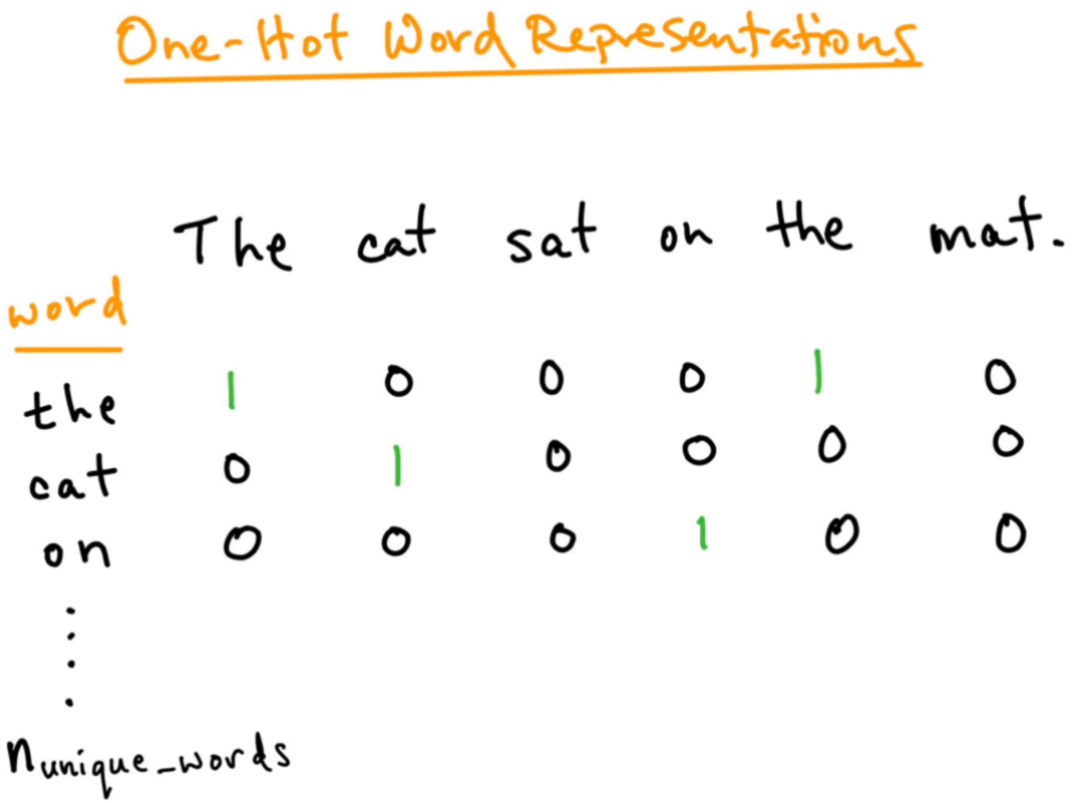
In the one-hot word representation, the sentence (the cat sat on the mat) is represented as a set of sparse vectors (mostly zeroes).



[TowardsDataScience1]

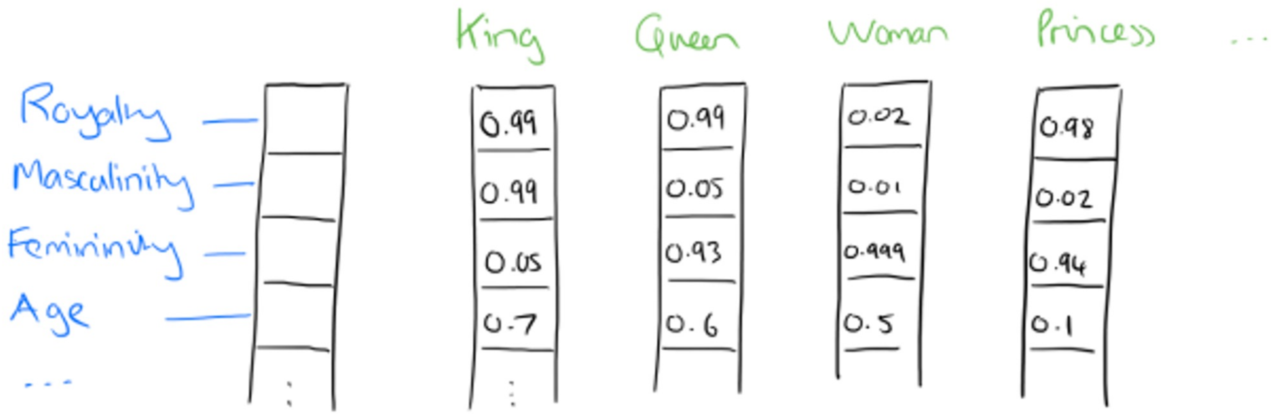
A small bite of Text Representations

This approach has two significant drawbacks: 1) **The huge memory capacity issues**, because of the sparse representation matrix; 2) **Lack of semantic understanding.** It can't understand **relationships** between words (e.g. school and book).



[TowardsDataScience1]

A small bite of Text Representations



In 2013, researchers from Google introduced a new model for text representation, which was revolutionary in NLP, named **word2vec** [Mikolov et al., 2013]. This shallow, deep learning model can represent words in dense vectors and capture semantic meaning between related terms (e.g. Paris and France, Madrid and Spain). Further research has built on top of **word2vec**, such as **GloVe** [Pennington et al., 2014] and **fastText** [Bojanowski et al., 2016].

Today this and similar representations are known as **word embeddings**.

[TowardsDataScience1]

A small bite of Text Representations, Transformer models

In late 2018, researchers from Google, again, came up with another model (**BERT**), which is considered the basis for state-of-the-art NLP research nowadays [Devlin et al., 2019], entirely based on the **Transformer** architecture [Vaswani et al., 2017].

12 giu 2017

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

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including

11 Oct 2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

A small bite of Text Representations, Transformer models

And then …

https://www.youtube.com/watch?v=K0cmmKPklp4

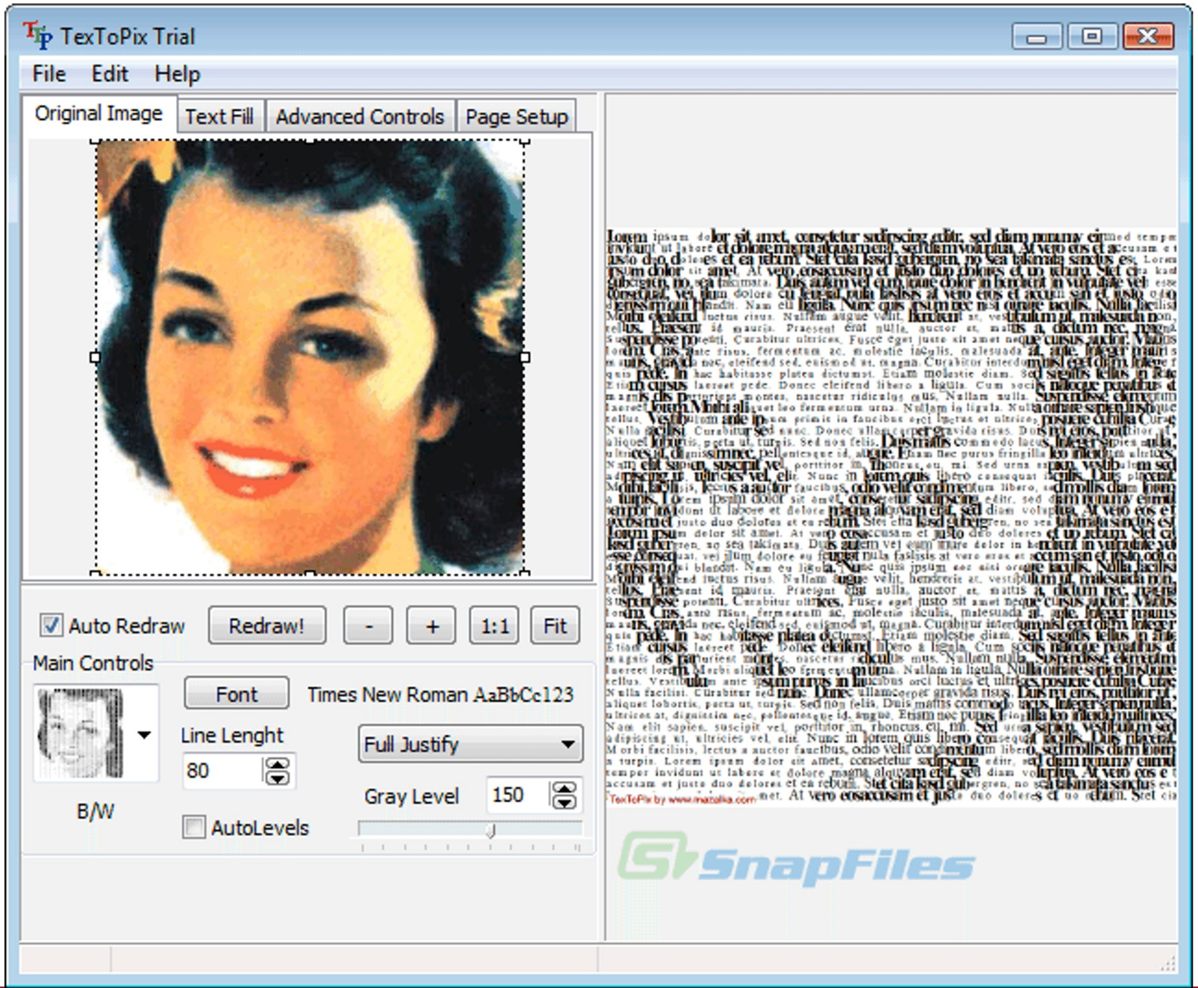
A small bite of Text Representations, Transformer models and Generative Pre-trained Transformer models

Generative NLP... OpenAI, ChatGPT (among others)

[TowardsDataScience1]

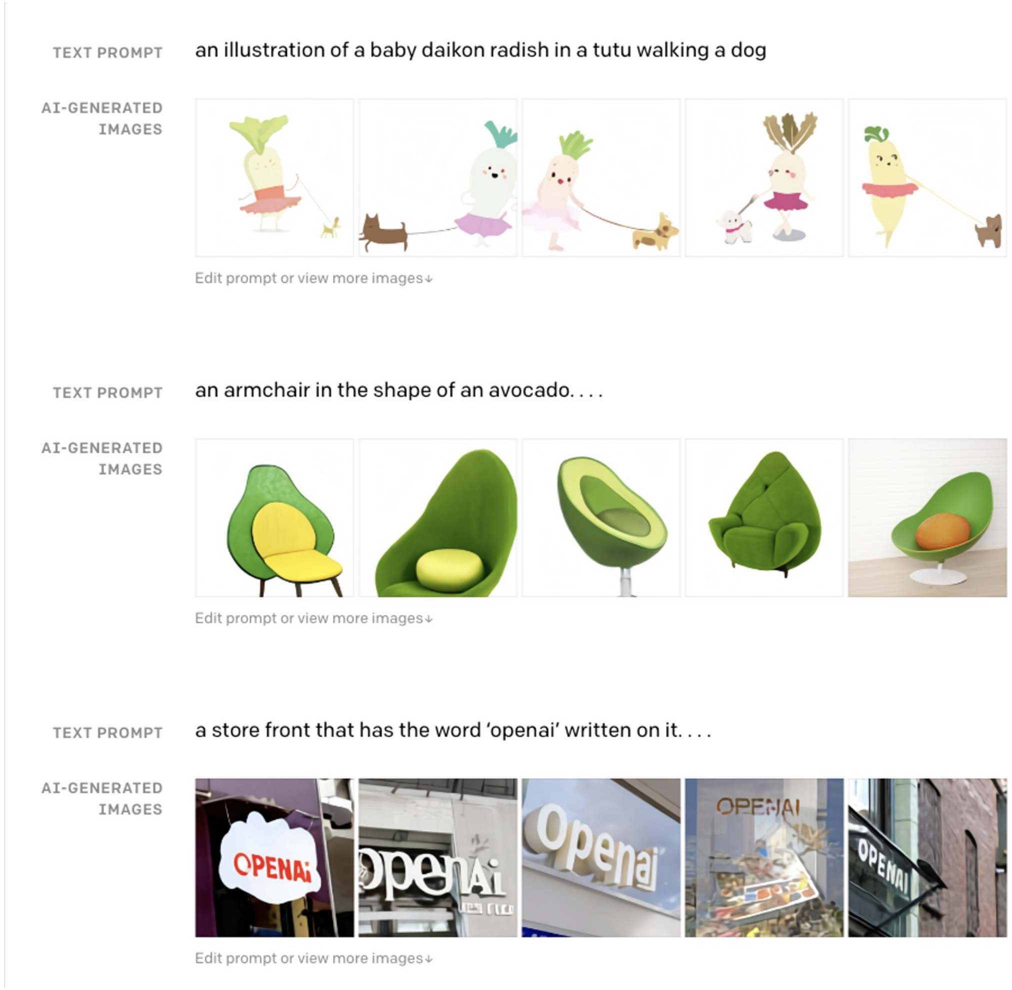
Multimodal NLP

Part II of the course also addresses the fascinating research field of Multimodal NLP



Multimodal NLP

Part II of the course also addresses the fascinating research field of Multimodal NLP



https://openai.com/blog/dall-e/

NLP Tasks and Research

Let's look at some NLP tasks and categorize them based on the research progress for the **English** language

[TowardsDataScience1]

NLP Tasks and Research

Let's look at some NLP tasks and categorize them based on the research progress for the **English** language

[TowardsDataScience1]

NLP Tasks and Research

"High Resource Languages vs Low Resource Languages"

"English is Neither Synonymous with Nor Representative of Natural Language"

The #BenderRule: On Naming the Languages We Study and Why It Matters [Bender, 2019]

[TowardsDataScience1]
[Bender, 2019]

NLP Tasks and Research

Let's look at some NLP tasks and categorize them based on the research progress for the **English** language

[TowardsDataScience1]

NLP Tasks and Research [ARRIVATO QUA CON FORMATTAZIONE E REFERENCES]

Let's look at some NLP tasks and categorize them based on the research progress for the **English** language

1. Mostly Solved:

- Text Classification *(e.g*. spam detection in Gmail).
- Part of Speech (**POS**) tagging: Given a sentence, determine the POS tag for each word (*e.g*. NOUN, VERB, ADV, ADJ).
- Named Entity Recognition (**NER**): Given a sentence, determine named entities (*e.g*. person names, locations, organizations).

[TowardsDataScience1]

NLP Tasks and Research

2. Making a Solid Progress:

- Sentiment Analysis: Given a sentence, determine it's polarity (*e.g*. positive, negative, neutral), or emotions (*e.g*. happy, sad, surprised, angry, etc)
- Co-reference Resolution: Given a sentence, determine which words ("mentions") refer to the same objects ("entities"). for example (**Manning** is a great NLP professor, **he** worked in the field for over two decades).
- Word Sense Disambiguation (**WSD**): Many words have more than one meaning; we have to select the meaning which makes the most sense based on the context (*e.g*. I went to the bank to get some money), here bank means a financial institution, not the land beside a river.
- Machine Translation (*e*.g. Google Translate).

[TowardsDataScience1]

3. Still Challenging:

- Dialogue agents and chat-bots, especially open-domain ones.
- Question Answering.
- Abstractive Summarization.
- NLP for low-resource languages (e.g. African languages, see Masakhane and *∀ et al., 2020*).

[TowardsDataScience1]

NLP Tasks and Research

3. New frontiers:

- … Image generation from texts.
- … Software requirements generation from texts.
- … Text generation from images, not simply optical characters recognition (OCR).
- ….
- …
- …
- …
- …
- …

[TowardsDataScience1]

NLP Tasks and Research

3. New frontiers:

- … Image generation from texts.
- … Software requirements generation from texts.
- … Text generation from images, not simply optical characters recognition (OCR).
- ….
- …
- …
- …
- …
- …

Can you name more examples?

DataScience1]

Resources and References

[TowardsDataScience1] Introduction to NLP: https://towardsdatascience.com/introduction-to-natural-language-processing-nlp-323cc007df3d

[Bender, 2013] Emily M. Bender, Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax, Synthesis Lectures on Human Language Technologies, June 2013, 184 pages, https://doi.org/10.2200/S00493ED1V01Y201303HLT020

[Mikolov et al., 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13). Curran Associates Inc., Red Hook, NY, USA, 3111–3119. https://dl.acm.org/doi/10.5555/2999792.2999959

[Pennington et al., 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. https://aclanthology.org/D14-1162/

[Bojanowski et al., 2016] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomás Mikolov: Enriching Word Vectors with Subword Information. CoRR abs/1607.04606 (2016) https://arxiv.org/abs/1607.04606

[Devlin et al., 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://aclanthology.org/N19-1423/

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need. CoRR abs/1706.03762 (2017) https://arxiv.org/abs/1706.03762

[JCR, 2020] Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B., AI WATCH. Defining Artificial Intelligence, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163. https://publications.jrc.ec.europa.eu/repository/handle/JRC118163

Resources and References

[Bender, 2019] Emily Bender, The #BenderRule: On Naming the Languages We Study and Why It Matters. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

[Jurafsky&Martin, 2022] Jurafsky and Martin. Speech and Language Processing, Prentice Hall, third edition https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Q&A

**Credits

The slides of this part of the course are the result of a personal reworking of the slides and of the course material from different sources:

1. The NLP course of Prof. Roberto Navigli, Sapienza University of Rome
2. The NLP course of Prof. Simone Paolo Ponzetto, University of Mannheim, Germany
3. The NLP course of Prof. Chris Biemann, University of Hamburg, Germany
4. The NLP course of Prof. Dan Jurafsky, Stanford University, USA