

Natural Language Processing with Deep Learning

RUHR
UNIVERSITÄT
BOCHUM

RUB

Lecture 1 — NLP tasks and evaluation

Prof. Dr. Ivan Habernal

October 16, 2025

www.trusthlt.org

Trustworthy Human Language Technologies Group (TrustHLT)

Ruhr University Bochum & Research Center Trustworthy Data Science and Security



CENTER FOR TRUSTWORTHY
DATA SCIENCE AND SECURITY

Motivation

- 1 Motivation
- 2 Course logistics
- 3 Challenges of NLP
- 4 Overview of typical NLP tasks
- 5 Evaluation

Why study deep learning for NLP?

Why study deep learning for NLP?

- 1 Build another AI startup and get filthy rich
- 2 Get the skill-set to destroy Skynet, if that happens

Why study deep learning for NLP?

- 1 Build another AI startup and get filthy rich
- 2 Get the skill-set to destroy Skynet, if that happens

But maybe also: Research on safety & security of LLMs, privacy, fairness, agentic AI, green AI, domain-specific LLMs, etc.

Preliminary course roadmap

- 1 NLP tasks and evaluation
- 2 Mathematical foundations of deep learning
- 3 Text classification 1: Log-linear models
- 4 Text classification 2: Deep neural networks
- 5 Text generation 1: LMs and word embeddings
- 6 Text classification 3: Encoding with RNNs
- 7 Text generation 2: Autoregressive RNNs and attention
- 8 Text classification 4: Self-attention and BERT
- 9 Text generation 3: Transformers
- 10 Text generation 4: Decoder-only models and GPT
- 11 Contemporary LLMs: Prompting and in-context learning
- 12 To be continued

Course logistics

- 1 Motivation
- 2 Course logistics**
- 3 Challenges of NLP
- 4 Overview of typical NLP tasks
- 5 Evaluation

Lecturers and tutors

Lecturers¹

- Prof. Dr. Ivan Habernal

Exercise tutors

- also me

¹`ivan.habernal@ruhr-uni-bochum.de`

Online resources

- Moodle for homework, announcements, and forum:
`https://moodle.ruhr-uni-bochum.de/course/view.php?id=66993`
- GitHub for lectures: `https://github.com/trusthlt/nlp-with-deep-learning-lectures`
- Discord as much faster forum: To be announced at Moodle
- Lectures recorded and published on YouTube

Textbooks and resources

- Recommended for each topic or lecture separately
- We'll use freely available resources (almost exclusively)

Top-notch research in NLP is "open source"

- Association for Computational Linguistics (ACL)
conferences in the "Anthology":
<https://aclanthology.org/>
- <https://arXiv.org>

Exercises

- Deepen your understanding of the matter, mostly hands-on
- Not graded
- We provide a guide-book and the assignment (mostly code, sometimes paper & pen)

Final exam

- February 11, 2026
- E-Assessment-Center in GAFO 04/402
- Exam questions: En
- Answers: En or De

It's your course, too!

Your feedback is very important

- Talk to me (live, discord, forum, e-mail)
- I'll post anonymous feedback forms regularly
- Slides issues: Just open bug/PR on GitHub

TrustHLT — Trustworthy Human Language Technologies

Research focus

- Privacy-preserving NLP (differential privacy; deep learning; representation learning; anonymization)
- Legal NLP, legal argumentation

Master thesis? HiWi job? Get in touch!

`ivan.habernal@ruhr-uni-bochum.de`

Challenges of NLP

- 1 Motivation
- 2 Course logistics
- 3 Challenges of NLP**
- 4 Overview of typical NLP tasks
- 5 Evaluation

Ambiguity and variability of human language

Example (Highly ambiguous)

Compare “I ate pizza with friends” to “I ate pizza with olives”

Example (Highly variable)

The core message of “I ate pizza with friends” can be expressed as “friends and I shared some pizza”

Y. Goldberg (2017). **Neural Network Methods for Natural Language Processing**. Morgan & Claypool

Humans — great users of language, very poor at formally understanding and describing rules that govern language

Language is challenging for machine learning (ML)

Natural language exhibits properties that make it very challenging for ML

- 1 Discrete
- 2 Compositional
- 3 Sparse

Language is symbolic and discrete

Basic elements of written language: **characters**

Characters form **words** that denote objects, concepts, events, actions, and ideas

Characters and words are discrete symbols

- Words such as “hamburger” or “pizza” each evoke in us a certain mental representations
- But they are distinct symbols, whose meaning is external to them, to be interpreted in our heads
- No inherent relation between “hamburger” and “pizza” can be inferred from the symbols or letters themselves

Characters and words are discrete symbols

Compare that to concepts such as **color** (in machine vision), or acoustic signals — these concepts are **continuous**

- Colorful image to gray-scale image using a simple mathematical operation
- We can compare two different colors based on inherent properties such as hue and intensity

This cannot be easily done with words

There is no simple operation to move from the word “red” to the word “pink” without using a large lookup table or a dictionary

Language is compositional

Letters → words → phrases → sentences

The meaning of a phrase can be larger than the meaning of the individual words, and follows a set of intricate rules

Example

Multi-word expressions (“New York”, “look something up”)

Idioms (“kick the bucket”, “blue chip”)

To interpret a text, we need to work beyond the level of letters and words, and look at long sequences of words such as sentences, or even complete documents

Data sparseness

Combinations of words to form meanings $\rightarrow \infty$

- We could never enumerate all possible valid sentences

No clear way of generalizing from one sentence to another, or defining the similarity between sentences, that does not depend on their meaning which is unobserved to us

Challenging when learning from examples

Even with a huge example set we are very likely to observe events that never occurred in the example set and that are very different

Overview of typical NLP tasks

- 1 Motivation
- 2 Course logistics
- 3 Challenges of NLP
- 4 Overview of typical NLP tasks**
- 5 Evaluation

Why are we learning this?

Important question to ask before we even start!

Deep learning is a tool and we need to understand

- Why we need this tool in the first place
- How do we know we have the right tool (it's doing its job well)

Coarse typology

Text classification and text generation

Overview of typical NLP tasks

Text classification tasks

Sentiment classification of movie reviews

Binary classification of reviews from IMDB

Example

Text: Read the book, forget the movie!

Label: Negative

→ semantic compositionality, long-range dependencies

- IMDB is the MNIST of NLP (the limus paper), 25k training, 25k test data points, balanced
- Why was it interesting?

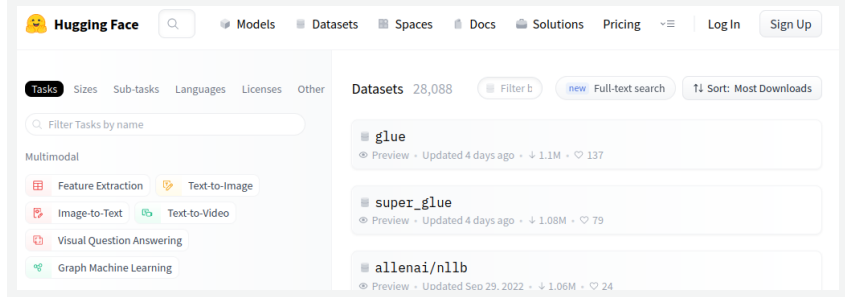
A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011).
“Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon: Association for Computational Linguistics, pp. 142–150

Task and dataset are used as synonyms

“The IMDB dataset” — must be properly cited! (incl. link)

Where to get datasets?

<https://huggingface.co/datasets>



The screenshot shows the Hugging Face website's 'Datasets' section. The top navigation bar includes the Hugging Face logo, a search bar, and links for Models, Datasets, Spaces, Docs, Solutions, Pricing, Log In, and Sign Up. The main content area is divided into two columns. The left column has tabs for Tasks, Sizes, Sub-tasks, Languages, Licenses, and Other. Below these is a search bar for tasks and a section for Multimodal tasks with buttons for Feature Extraction, Text-to-Image, Image-to-Text, Text-to-Video, Visual Question Answering, and Graph Machine Learning. The right column displays a list of datasets. The first dataset is 'glue', updated 4 days ago, with 1.1M downloads and 137 likes. The second is 'super_glue', also updated 4 days ago, with 1.08M downloads and 79 likes. The third is 'allenai/nllb', updated Sep 29, 2022, with 1.06M downloads and 24 likes. Each dataset entry includes a 'Preview' link.

Natural Language Inference

Two sentences: entailment, contradiction, or neutral?

Example

Text: A soccer game with multiple males playing.

Hypothesis: Some men are playing sport.

Label: Entailment

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning (2015). “**A large annotated corpus for learning natural language inference**”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642

The “standard” NLI paper and dataset from Stanford: SNLI

- 570k human-written English sentence pairs

How is SNLI data different from the IMDB?

- IMDB data that was “annotated for free” by each author

Side step 1: Gold standard data

- Many datasets are annotated by experts, super costly
- Each example by multiple annotators, then the final “gold” label is decided upon

How to measure task subjectivity and annotation quality?

Inter-Annotator Agreement

Take chance agreement into account

- Cohen's Kappa, Scott's Pi, Krippendorff's Alpha, Krippendorff's Unitized Alpha (Artstein and Poesio, 2008)

I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. Spiecker genannt Döhmann, and C. Burchard (2023). **“Mining Legal Arguments in Court Decisions”**. In: *Artificial Intelligence and Law*

R. Artstein and M. Poesio (2008). **“Inter-Coder Agreement for Computational Linguistics”**. In: *Computational Linguistics* 34.4, pp. 555–596

Side step 2: Who creates these tasks and why?

- Mostly researchers
- Mostly for phenomena in language and to which extent NLP can “solve” them
- Shared datasets became popular with machine learning in NLP

Tasks are classified into various (arbitrary) taxonomies with (mostly agreed upon) names, for example

- Sentiment analysis \in text classification
- SNLI \in sentence-pair classification

Deeper in sentences: NER

Named entity recognition: Find entities of predefined types

Example

U.N.	Organization
official	
Ekeus	Person
heads	
for	
Baghdad	Location
.	

E. F. Tjong Kim Sang and F. De Meulder (2003). **"Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition"**. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. <https://aclanthology.org/W03-0419>, pp. 142–147

How to model and annotate such a task?

NER: Sequence labeling task

Tokenize, assign each word a type

Example

U.N.	I-ORG
official	O
Ekeus	I-PER
heads	O
for	O
Baghdad	I-LOC
.	O

E. F. Tjong Kim Sang and F. De Meulder (2003). **"Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition"**. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. <https://aclanthology.org/W03-0419>, pp. 142–147

CoNLL 2003: Four entities (PER, ORG, LOC, MISC)

NER: BIO encoding

What if two consequent tokens are same type?

“Whenever two entities of type XXX are immediately next to each other, the first word of the second entity will be tagged B-XXX in order to show that it starts another entity”

BIO encoding

An instance of Multi-class classification on token level

E. F. Tjong Kim Sang and F. De Meulder (2003). **“Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition”**. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. <https://aclanthology.org/W03-0419>, pp. 142–147

SuperGLUE

SuperGLUE — popular benchmark collection of various tasks/datasets in English

*“The goal of SuperGLUE is to provide a simple, robust evaluation metric of any method capable of being applied to a broad range of **language understanding tasks**.”*

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman (2019). **“SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”**. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, Inc., pp. 3266–3280

Extractive Question Answering: SQuAD 2.0

Example

Endangered Species Act Paragraph: "... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a **1937 treaty** prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little **opposition** was raised."

Question 1: "Which laws faced significant **opposition**?"

Plausible Answer: later laws

Question 2: "What was the name of the **1937 treaty**?"

Plausible Answer: Bald Eagle Protection Act

P. Rajpurkar, R. Jia, and P. Liang (2018).
"Know What You Don't Know: Unanswerable Questions for SQuAD". In:
Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, pp. 784–789

Unanswerable questions w/ plausible (but incorrect) answers. Relevant keywords are **bold**.

Overview of typical NLP tasks

Text generation tasks

Machine translation

<u>NEWS OF THE MONTH</u>	
	 Humus
• Gnocchi Sorrento style	€10,50
• Soup of the day with vegetables and rice	€10,00
• Mixed grill	€28,00
with potatoes and salad (x2 people)	
• White pizza	€10.50
with potatoes, mushrooms, and sausage	
<u>NACHRICHTEN DES MONATS</u>	
• Gnocchi Sorrentinischer Art	10,50 €
• Tagessuppe mit Gemüse und Reis	10,00 €
	28,00 €

O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, P. Koehn, and C. Monz (2018). **"Findings of the 2018 Conference on Machine Translation (WMT18)"**. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Vol. 2. Brussels, Belgium: Association for Computational Linguistics, pp. 272–303

Standard datasets from WMT (formerly Workshop on MT)

Machine translation

Figure 1.1 Ten translators translate the same short French sentence—*Sans se démonter, il s’est montré concis et précis.*—in 10 different ways. Human evaluators also disagree for each translation if it is correct or wrong.

Assessment Correct/Wrong	Translation
1/3	<i>Without fail, he has been concise and accurate.</i>
4/0	<i>Without getting flustered, he showed himself to be concise and precise.</i>
4/0	<i>Without falling apart, he has shown himself to be concise and accurate.</i>
1/3	<i>Unswayable, he has shown himself to be concise and to the point.</i>
0/4	<i>Without showing off, he showed himself to be concise and precise.</i>
1/3	<i>Without dismantling himself, he presented himself consistent and precise.</i>
2/2	<i>He showed himself concise and precise.</i>
3/1	<i>Nothing daunted, he has been concise and accurate.</i>
3/1	<i>Without losing face, he remained focused and specific.</i>
3/1	<i>Without becoming flustered, he showed himself concise and precise.</i>

Source: P. Koehn (2020). **Neural Machine Translation.** (not freely available). Cambridge University Press

(Abstractive) Document summarization

K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom (2015). **“Teaching Machines to Read and Comprehend”**.

In: *Proceedings of NeurIPS*. Curran Associates, Inc., pp. 1–9

Popular dataset: CNN/Daily Mail

- Online news articles (781 tokens on average)
- Paired with multi-sentence summaries (3.75 sentences or 56 tokens on average)
- 287k training pairs, 13k validation pairs, 11k test pairs

Dialogue: PersonaChat

165k utterances; Task: next utterance prediction

Persona 1	Persona 2
I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones

[PERSON 1:] Hi

[PERSON 2:] Hello ! How are you today ?

[PERSON 1:] I am good thank you , how are you.

[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.

[PERSON 1:] Nice ! How old are your children?

[PERSON 2:] I have four that range in age from 10 to 21. You?

[PERSON 1:] I do not have children at the moment.

[PERSON 2:] That just means you get to keep all the popcorn for yourself.

[PERSON 1:] And Cheetos at the moment!

[PERSON 2:] Good choice. Do you watch Game of Thrones?

[PERSON 1:] No, I do not have much time for TV.

[PERSON 2:] I usually spend my time painting: but, I love the show.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston (2018).
“Personalizing Dialogue Agents: I have a dog, do you have pets too?” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2204–2213

Overview of some typical generation tasks

A. B. Sai, A. K. Mohankumar, and M. M. Khapra (2023). “A Survey of Evaluation Metrics Used for NLG Systems”. In: *ACM Computing Surveys* 55.2, pp. 1–39

NLG task	Context (Input)	Reference
Machine Translation (MT)	Source language sentence	Translation
Abstractive Summarization (AS)	Document	Summary
Question Answering (QA)	Question + Background info (Passage, Image, etc)	Answer
Question Generation (QG)	Passage, Knowledge base, Image	Question
Dialogue Generation (DG)	Conversation history	Response
Image Captioning (IC)	Image	Caption
Data to Text (D2T)	Semi-structured data (Tables, Graphs)	Description

Table 1: Context and references for NLG tasks

Overview of typical NLP tasks

Classification as generation

Unifying classification and generation

Any task incl. classification → "text-to-text" format

Example (Translation En-De)

Input: *translate English to German: That is good.*

Expected output text: *Das ist gut.*

Example (MNLI)

Input: *mnli premise: I hate pigeons. hypothesis: My feelings towards pigeons are filled with animosity.* Expected output text: *entailment*

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). **"Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"**. In: *Journal of Machine Learning Research* 21.140, pp. 1–67

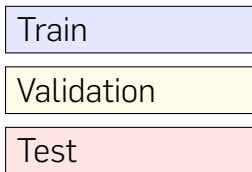
Evaluation

- 1 Motivation
- 2 Course logistics
- 3 Challenges of NLP
- 4 Overview of typical NLP tasks
- 5 Evaluation**

Train/Dev/Test data splits

Training and Test data

Development (Validation) set used for optimizing hyper-parameters



Evaluation

Evaluation of text classification

Confusion matrix (binary case)

Two classes: Positive and Negative

Confusion matrix

	Predicted Neg	Predicted Pos
Actually Neg	True negative (TN)	False positive (FP)
Actually Pos	False negative (FN)	True positive (TP)

Ordering of columns and rows is **arbitrary**!

Accuracy

Accuracy of classifier f on test set T :

$$\text{Acc}_T(f) = \frac{1}{|T|} \sum_{i=1}^{|T|} I(f(x_i) = y_i)$$

N. Japkowicz and M. Shah (2011). **Evaluating Learning Algorithms: A Classification Perspective.** (not freely available). Cambridge University Press

Example (Disease detection)

	Pred. Negative	Pred. Positive
Act. Negative	168	33
Act. Positive	48	37

$37 + 48 + 33 + 168 = 286 \rightarrow$ Test set size $|T| = 286$

$$\text{Acc}_T(f) = \frac{1}{286} (37 + 168) = 0.7186$$

Precision, recall, F-1 score

Confusion matrix

	Pred. Negative	Pred. Positive
Act. Negative	True negative (TN)	False positive (FP)
Act. Positive	False negative (FN)	True positive (TP)

Precision (for class positive) = $TP / (TP + FP)$

Recall (for class positive) = $TP / (TP + FN)$

F-1 score (for class positive) = $2PR / (P + R)$

Confusion matrix – multi-class

true class:	prediction:	<i>money-fx</i>	<i>trade</i>	<i>interest</i>	<i>wheat</i>	<i>corn</i>	<i>grain</i>
<i>money-fx</i>		95	0	10	0	0	0
<i>trade</i>		1	1	90	0	1	0
<i>interest</i>		13	0	0	0	0	0
<i>wheat</i>		0	0	1	34	3	7
<i>corn</i>		1	0	2	13	26	5
<i>grain</i>		0	0	2	14	5	10

Confusion matrix — multi-class

- We can unambiguously compute Precision and Recall for each class
- How to get the F-1 score for the complete test set across classes?
 - Macro-averaging (average of F-1 scores), or micro-averaging
 - These details might get tricky so always report exactly what you do!

M. Sokolova and G. Lapalme (2009). **"A systematic analysis of performance measures for classification tasks"**. In: *Information Processing and Management* 45.4, pp. 427–437

Evaluation

Evaluation of text generation

Evaluating text generation is hard

A. B. Sai, A. K. Mohankumar, and M. M. Khapra (2023). "A Survey of Evaluation Metrics Used for NLG Systems". In: *ACM Computing Surveys* 55.2, pp. 1–39

Table 3. Automatic Metrics That have been Proposed (✓) or Adopted (*) for Various NLG Tasks

Metric	Tasks the metric is proposed or adopted for:										Resources used (at run/test time)
	MT	AS	DG	IC	QA	D2T	QG	≥ 0	IoI	sym	
Context-free metrics											
BLEU [94]	✓	*	*	*	*	*	*	✓	✓		tokenizer
NIST [34]	✓	*	*	*	*	*	*	✓	✓		tokenizer
METEOR [7]	✓	*	*	*	*	*	*	✓			tokenizer, WordNet, stemmer
ROUGE [70]	*	✓	*	*	*	*	*	✓			tokenizer
GTM [132]	✓	*	*	*	*			✓	✓		tokenizer
CIDEr [135]				✓				✓			tokenizer
SPICE [5]				✓				✓			tokenizer, stemmer, word frequencies (TF-IDF)
SPIDer [72]				✓				✓			SPICE, CIDEr
WER	*							✓	✓		tokenizer
MultiWER	✓							✓	✓		tokenizer
TER [122]	✓							✓	✓		tokenizer
ITER [93]	✓							✓	✓		tokenizer
CDER [64]	✓							✓	✓		tokenizer
chrF [100]	✓	*		*				✓	✓		–
characTER [138]	✓							✓	✓		tokenizer
EED [123]	✓							✓	✓		tokenizer
Vector Extrema [42]	*	*	*	*	*	*			✓		tokenizer, pretrained embeddings
Vector Autoregression [40]	*	*	*	*	*	*			✓		tokenizer, pretrained embeddings

BLEU (Bilingual Evaluation Understudy)

Almost first and most popular metric for MT

- Precision-based metric that computes the n-gram overlap between the reference and the hypothesis
- In particular, BLEU is the ratio of the number of overlapping n-grams to the total number of n-grams in the hypothesis.

Corpus-level metric, i.e., BLEU gives a score over the entire corpus (as opposed to scoring individual sentences)

Major drawbacks of BLEU: (i) it does not take recall into account and (ii) it only allows exact n-gram matching

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu (2002). **"BLEU: a Method for Automatic Evaluation of Machine Translation"**. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: Association for Computational Linguistics, pp. 311–318

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE metric includes a set of variants: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S

- ROUGE-N is similar to BLEU-N in counting the n-gram matches between the hypothesis and reference, however, it is a recall-based measure unlike BLEU which is precision-based
- ROUGE-L measures the longest common subsequence (LCS) between a pair of sentences

C.-Y. Lin (2004). **"ROUGE: A Package for Automatic Evaluation of Summaries"**. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81

Evaluation

Caveats of NLP benchmarking

Is BLEU a good metric?

BLEU was originally proposed for diagnostic evaluations of MT systems, that is, as a technique for allowing researchers and developers to quickly “weed out bad ideas from good ideas”

- Wide range of correlations between BLEU and human evaluations
- BLEU should not be the primary evaluation technique in NLP papers

E. Reiter (2018). **“A Structured Review of the Validity of BLEU”**. In: *Computational Linguistics* 44.3, pp. 393–401

The ‘gold’ data paradigm might not always fit

The assumption of a ground truth makes sense when humans highly agree on the answer

- “Does this image contain a bird?”
- “Is ‘learn’ a verb?”
- “What is the capital of Italy?”

This assumption often does not make sense, especially when language is involved

- “Is this comment toxic?”

Human label variation impacts all steps of the traditional ML pipeline, and is an opportunity, not a problem

B. Plank (2022). “**The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation**”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10671–10682

New and new benchmarks...

Ecology

Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

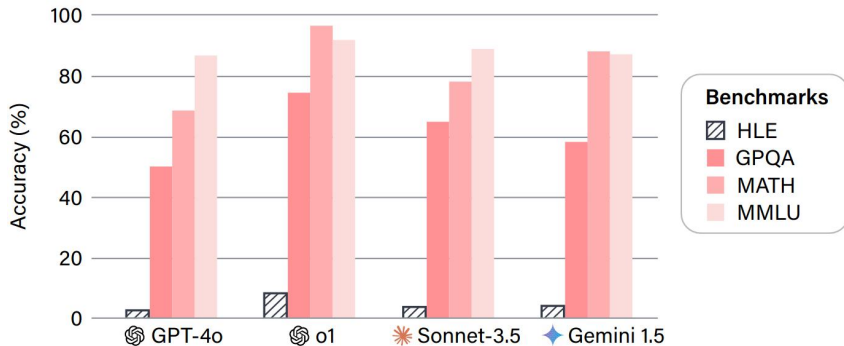
Figure 1: Example from the Humanity's Last Exam dataset. The benchmark contains 2,500 challenging questions across over a hundred subjects.

Phan et al. (Sept. 2025). **"Humanity's Last Exam"**. In: *arXiv preprint*

But performance of LLMs keeps rising

Phan et al. (Sept. 2025). **"Humanity's Last Exam"**. In: *arXiv preprint*

Accuracy of LLMs Across Benchmarks



Beware of (unintended) data contamination!

Data contamination = evaluate LLMs on the same data they were trained on

“90 papers accessed ChatGPT through the web interface, hence providing data that OpenAI could have used to further improve its models”

S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek (2024). **“Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs”**. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. St. Julian's, Malta: Association for Computational Linguistics, pp. 67–93

Recap

- 1 Motivation
- 2 Course logistics
- 3 Challenges of NLP
- 4 Overview of typical NLP tasks
- 5 Evaluation

Takeaways: We set up the scene

- NLP is challenging
- Vast amount of tasks and datasets
- Data quality matters
- Understanding the data, annotators, task matters too
- Deep familiarity with common evaluation metrics is essential
- Getting better scores is just a beginning of the story
- Evaluating generation is an art

License and credits

Licensed under Creative Commons
Attribution-ShareAlike 4.0 International
(CC BY-SA 4.0)



Credits

Ivan Habernal

Content from ACL Anthology papers licensed under CC-BY
<https://www.aclweb.org/anthology>