Natural Language Processing with Deep Learning

Lecture 1 — NLP tasks and evaluation

Prof. Dr. Ivan Habernal

October 13, 2023

Natural Language Processing Group
Paderborn University
We focus on Trustworthy Human Language Technologies

www.trusthlt.org

Motivation

Why study deep learning for NLP?

1. GPT-4

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

Lecturers and tutors

Lecturers[1]

- Prof. Dr. Ivan Habernal

Tutors

- also me

[1] ivan.habernal@uni-paderborn.de

Online resources

- PANDA for homework, announcements, and forum: `TODO`
- GitHub for lectures: `https://github.com/trusthlt/nlp-with-deep-learning-lectures`
- Discord as much faster forum: `https://discord.gg/QAySNRUUJ`
- 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 and Homeworks

Exercises (EX)

- Deepen your understanding of the matter
- Not graded

Homeworks (HW)

- Get hands dirty
- Submit in groups of two
- $\geq 70\%$ of HW points $\rightarrow$ eligible for 0.3/0.4 exam bonus
- **Follow announcements and instructions on Moodle**

Final exam

PADERBORN UNIVERSITY

- Date to be announced
- Exam questions: En
- Answers: En or De

It’s your course, too!

Your feedback is very important

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

Trustworthy Human Language Technologies

PADERBORN UNIVERSITY

Research focus

- Privacy-preserving NLP (differential privacy; deep learning; representation learning; graph networks)
- Argument mining "that matters" (legal argument mining; ethical argumentation)

Master thesis? HiWi job? Get in touch!

`ivan.habernal@uni-paderborn.de`

Challenges of NLP

Ambiguity and variability of human language

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

Highly ambiguous

Example

Compare "I ate pizza with friends" to "I ate pizza with olives"

Highly variable

Example

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

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

Supervised machine learning to save us...

The best known set of methods for dealing with language data $\rightarrow$ supervised machine learning algorithms

- ML attempts to infer patterns and regularities from a set of pre-annotated input–output pairs
- ML excels at problem domains where a good set of rules is very hard to define but annotating the expected output for a given input is relatively simple

...but language is even more challenging

Natural language exhibits properties that make it even more 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 $\rightarrow$ words $\rightarrow$ phrases $\rightarrow$ 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

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

$\rightarrow$ 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?

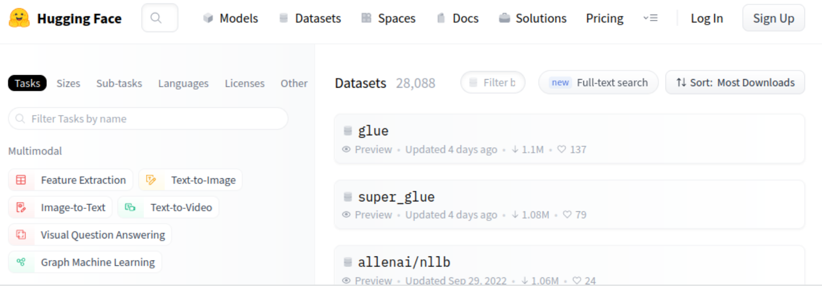
A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011). **"Learning Word Vectors for Sentiment Analysis".** In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon: Association for Computational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



Natural Language Inference

PADERBORN UNIVERSITY

Two sentences: entailment, contradiction, or neutral?

Example

Text: A soccer game with multiple males playing.
Hypothesis: Some men are playing sport.
Label: Entailment

The "standard" NLI paper and dataset from Stanford: SNLI

- 570k human-written English sentence pairs
- manually labeled for balanced classification

How is SNLI data different from the IMDB?

- IMDB data that was "annotated for free" by each author

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

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
Baghdat       Location
.
```

How to model and annotate such a task?

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

NER: Sequence labeling task

Tokenize, assign each word a type

Example

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

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

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

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

Recognizing Textual Entailment (RTE)

I. Dagan, B. Dolan, B. Magnini, and D. Roth (2009). **“Recognizing textual entailment: Rational, evaluation and approaches”.** In: *Natural Language Engineering* 15.4, pp. 1–27

Two-class (binary) classification

Whether or not the Text entails the Hypothesis

Example

Text: Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.

Hypothesis: Christopher Reeve had an accident.

Entailment: False

Note: SNLI adapted RTE!

Coreference resolution (WSC — Winograd Schema Challenge)

H. J. Levesque, E. Davis, and L. Morgenstern (2012). **"The Winograd Schema Challenge".** In: *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. Rome, Italy: Association for the Advancement of Artificial Intelligence, pp. 552–561

Examples consist of a sentence with a pronoun and a list of noun phrases from the sentence

Determine the correct referrent of the pronoun

Example

Text: Mark told <u>Pete</u> many lies about himself, which Pete included in his book. <u>He</u> should have been more truthful.

Coreference: False

$\rightarrow$ everyday knowledge and commonsense reasoning to solve

BoolQ

Each example: short passage and a yes/no question about the passage

Example

Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A: Yes. [An example event is given.]

$\rightarrow$ complex, non-factoid information, requires difficult entailment-like inference to solve

C. Clark, K. Lee, M.-w. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova (2019). **"BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions".** 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2924–2936

MultiRC: Multi-Sentence Reading Comprehension

D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth (2018). **"Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences".** In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, LA: Association for Computational Linguistics, pp. 252–262

Each example consists of

- Context paragraph
- Question about that paragraph
- List of possible answers (true/false)

Desirable properties:

1. Multiple possible correct answers $\rightarrow$ each question-answer pair must be evaluated independent of other pairs
2. Answering each question requires drawing facts from multiple context sentences

Extractive Question Answering: SQuAD 2.0

PADERBORN UNIVERSITY

Example

Endangered Species Act Paragraph: "... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a **1937 treaty** prohibiting the hunting ofright and gray whales, and the <u>Bald Eagle Protection Act of 1940</u>. These <u>later laws</u> 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: <u>later laws</u>

Question 2: "What was the name ofthe **1937 treaty**?"
Plausible Answer: <u>Bald Eagle Protection Act</u>

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

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

Machine translation is still hard! (Feb 2023)

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

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

(Abstractive) Document summarization

PADERBORN UNIVERSITY

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

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

165k utterances; Task: next utterance prediction

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

Overview of typical NLP tasks

Classification as generation

Unifying classification and generation

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

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*

Evaluation

Train/Dev/Test data splits

Training and Test data

Development (Validation) set used for optimizing hyper-parameters

Train

Validation

Test

Cross validation

K-fold cross-validation partitions the data into K chunks

$K - 1$ of which form the training set $\mathcal{R}$

The last chunk serves as the test set $\mathcal{V}$ (or validation)

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	Train	Train	Train	Test
Train	Train	Train	Test	Train
Train	Train	Test	Train	Train
Train	Test	Train	Train	Train
Test	Train	Train	Train	Train

Figure 1: Example of 5-fold CV

Evaluation

Evaluation of text classification

Confusion matrix (binary case)

Two classes: Positive and Negative

Confusion matrix

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

Act. Negative = Actually negative = Gold label

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

Accuracy

PADERBORN UNIVERSITY

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

Accuracy of classifier f on test set T:

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

Example (Disease detection)

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

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

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

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:	*money-fx*	*trade*	*interest*	*wheat*	*corn*	*grain*
money-fx	95	0	10	0	0	0
trade	1	1	90	0	1	0
interest	13	0	0	0	0	0
wheat	0	0	1	34	3	7
corn	1	0	2	13	26	5
grain	0	0	2	14	5	10

Confusion matrix – multi-class

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

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!

Evaluation

Evaluation of text generation

More text 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

Table 2. Context and Reference/Hypothesis Forms for Each NLG Task

NLG task	Context (Input)	Reference and Hypothesis
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, AMRs, *etc*)	Description

Evaluating text generation is hard

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

Metric	Tasks the metric is proposed or adopted for: MT	AS	DG	IC	QA	D2T	QG	≥ 0	IoI	sym	Resources used (at run/test time)
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 Averaging [63]	*	*	*	*	*	*			✓		tokenizer, pretrained embeddings
Greedy matching [107]	*	*	*	*	*	*			✓		tokenizer, pretrained embeddings
WMD [62]	*	*		*				✓	✓	✓	tokenizer, BOW vectors, pretrained embeddings
WEWPI [37]	*	*		*				✓	✓	✓	tokenizer, BOW vectors, pretrained embeddings
MEANT [76]	✓	*						✓	✓	✓	tokenizer, BOW vectors, pretrained embeddings
YiSi [74]	✓							✓	✓	✓	tokenizer, BOW vectors, pretrained embeddings
BERTr [84]	*								✓		tokenizer, BERT embeddings
BERTscore [144]	✓		*	✓	*			✓	✓	✓	tokenizer, BERT embeddings
MoverScore [147]	✓	✓		✓		✓			✓		tokenizer, contextualized embeddings
BEER [125]	✓								✓		statistical features (eg: F1), permutation trees

BLEU (Bilingual Evaluation Understudy)

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

Among the first and most popular metrics proposed for automatic evaluation of MT systems

- 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

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

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

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

Evaluation

Caveats of NLP benchmarking

The ‘gold’ data paradigm might not always fit

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

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

- Questions determining a word sense
- “Is this comment toxic?”

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

Human annotators are biased

PADERBORN UNIVERSITY

M. Geva, Y. Goldberg, and J. Berant (2019). **"Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets".** In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1161–1166

Datasets are often constructed using a small number of annotators, and humans are biased

- Concerns about data diversity, especially when workers freely generate sentences
- Models do not generalize well to examples from annotators that did not contribute to the training set

Artifacts in datasets

Datasets have artifacts (spurious statistics) that can be exploited

Claim	Google is not a harmful monopoly
Reason	People can choose not to use Google
Warrant	Other search engines don't redirect to Google
Alternative	All other search engines redirect to Google

Reason (and since) **Warrant** $\rightarrow$ **Claim**
Reason (but since) **Alternative** $\rightarrow$ $\neg$ **Claim**

Figure 1: An example of a data point from the ARCT test set and how it should be read. The inference from R and A to $\neg C$ is by design.

I. **Habernal**, H. Wachsmuth, I. Gurevych, and B. Stein (2018). **"The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants".** In: *Proceedings of NAACL*. New Orleans, LA, pp. 1930–1940

T. Niven and H.-Y. Kao (2019). **"Probing Neural Network Comprehension of Natural Language Arguments".** In: *Proceedings of ACL*. Florence, Italy, pp. 4658–4664

Recap

Motivation

Course logistics

Challenges of NLP

Overview of typical NLP tasks

- Text classification tasks
- Text generation tasks
- Classification as generation

Evaluation

- Evaluation of text classification
- Evaluation of text generation
- Caveats of NLP benchmarking

Takeaways: We set up the scene

PADERBORN UNIVERSITY

- 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

Credits

Ivan Habernal