

Natural Language Processing: Introduction

H02B1A



Miryam de Lhoneux

*Department of Computer Science
KU Leuven*

Natural language processing

The screenshot shows a Google search for 'natural language processing'. The search bar at the top contains the text 'natural language processing' with a close button (X) and icons for voice search, image search, and general search. Below the search bar, tabs for 'All', 'Images', 'Videos', 'News', 'Books', 'Web', and 'Finance' are visible, with 'All' selected. The main search result is a snippet: 'Natural language processing (NLP) is the ability of a computer program to understand human language'. To the right of this snippet, a red box highlights the word 'language' in the title 'Natural language processing', and a tooltip appears with the text 'Correct your spelling language'. Below the search results, there is a language selection bar with 'Detect language', 'Dutch', 'English' (selected), and 'French'. Below this, two panels are shown: the left panel contains the text 'natural language processing' with a close button (X), a microphone icon, a speaker icon, and a character count '27 / 5,000'; the right panel contains the Dutch translation 'natuurlijke taalverwerking' with a star icon, a microphone icon, and share icons (copy, link, share).

Natural language processing: definition

Natural language processing is a field of science and engineering focused on the development and study of **automatic systems** that **understand** and **generate** natural (that is, human,) languages.

StanfordNLP (2024)

Natural language processing: all-in-one

Give me a short definition of Natural Language Processing, then translate it to Dutch. But first, correct my spelling.



Corrected Spelling:

- **Original:** "Natural Langaue Processing"
- **Corrected:** "Natural **Language** Processing"

Short Definition of Natural Language Processing:

Natural Language Processing (NLP) is a field of AI that enables computers to understand, analyze, and generate human language.

Translation to Dutch:

Natural Language Processing (NLP) is een tak van AI die computers in staat stelt om menselijke taal te begrijpen, analyseren en genereren.

Retrieved 19/08/2024

Natural language processing: all-in-one

Who teaches NLP in the north pole?



In the North Pole, you'd probably find Santa Claus teaching NLP (Natural Language Processing) to his elves! With all that magical efficiency and gift production, Santa's elves would need to be well-versed in processing language and optimizing communication to manage all the holiday requests. Plus, with Santa's knack for understanding what people want, he'd make a great NLP instructor!

Retrieved 19/08/2024

This course

Foundations

- Building blocks of state-of-the-art NLP systems
- What challenges have been overcome?
- What challenges are still open?

Tasks and applications

- Tasks that cover all levels of linguistic processing (morphology, syntax, semantics)
- Example applications (e.g. machine translation and conversational agents)

This course

Learning goals

At the end of the course, you should be able to

- describe fundamental concepts and algorithms seen in class
- implement and apply these algorithms
- describe and use state-of-the-art NLP models

This lecture

- Current state of NLP
- General challenges
- Representing words and sentences
- Practical aspects of the course

State of NLP

mostly solved

Spam detection

Let's go to Agra!

Buy V1AGRA ...



Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party,
Friday May 27 at 8:30

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing
fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is
good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you
want a ticket?

NLP progress in 2012 by Dan Jurafsky

State of NLP

mostly solved

Spam detection

Let's go to Agra!

Buy VIAGRA ...



Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Parsing

I can see Alcatraz from the window!

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.



Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my mouse.



Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party,
Friday May 27 at 8:30making good progress
~~still really hard~~ but...

Question answering (QA)

Q. How effective is ibuprofen in reducing
fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is
good

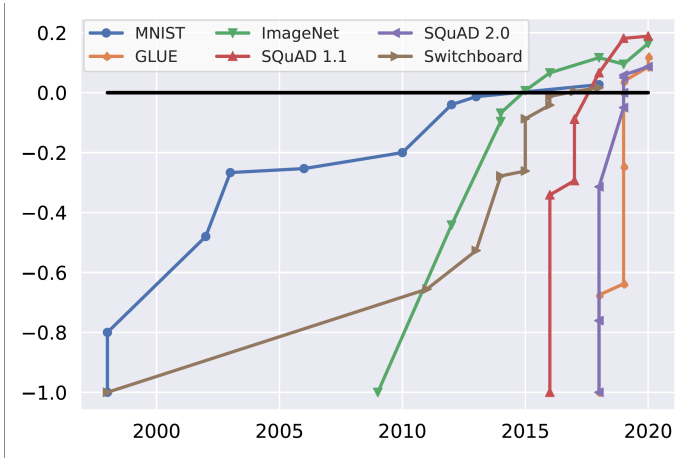
Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you
want a ticket?

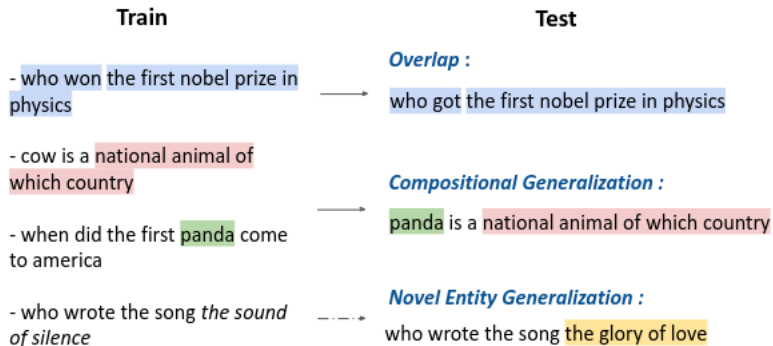
NLP progress in 2024, adapted from Dan Jurafsky

State of NLP



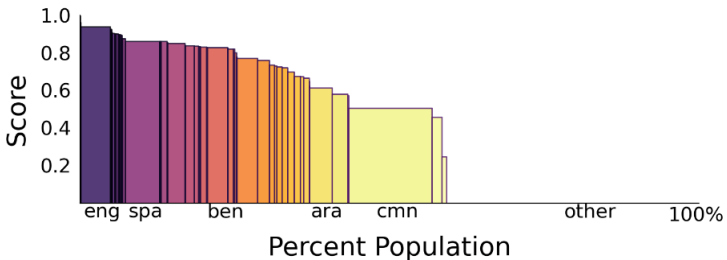
Benchmark progress in NLP and CV. -1 =initial performance, 0 =human performance.
Kielbaso et al. (2021).

State of NLP



Challenges of open-domain QA (Liu et al., 2022).

State of NLP

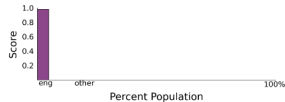


Performance disparity of NER systems considering languages and their number of speakers (Song et al., 2023).

State of NLP



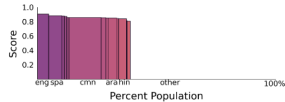
(a) KG Link Tail Prediction



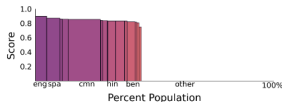
(b) Text Classification



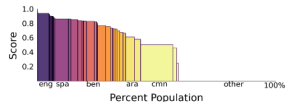
(c) Machine Translation



(d) Text Pair Classification



(e) Extractive QA



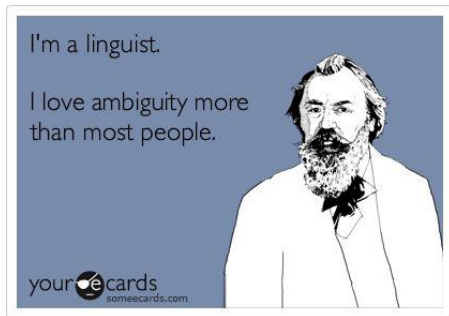
(f) Named Entity Recognition

Performance disparity of systems considering languages and their number of speakers for 6 tasks. (Song et al., 2023).

What makes NLP challenging?

- Ambiguity
- Non-standard language
- Neologisms
- Need for external knowledge
- Long tail of rare words/phenomena

Ambiguity



- I love ambiguity more than most people love ambiguity.
- I love ambiguity more than I love most people.

Ambiguity

I made her duck

- I cooked duck for her
- I cooked a duck which belonged to her
- I created the duck which belongs to her
- I caused her to lower her head
- I turned her into a duck

Need context to disambiguate.

Ambiguity

Types of ambiguity

- Lexical: *I saw a bat*
 saw = past tense of see or present tense of saw
 bat = flying mammal or wooden club
- Structural: *I saw a man with a telescope*
 I have a telescope or *the man* does
- Semantic: *John and Mary are married*
 to each other or to other people
- Anaphoric: *Margaret invited Susan, and she gave her a sandwich.*
 she = Margaret or Susan

Exercise: find all the ambiguities in this sentence

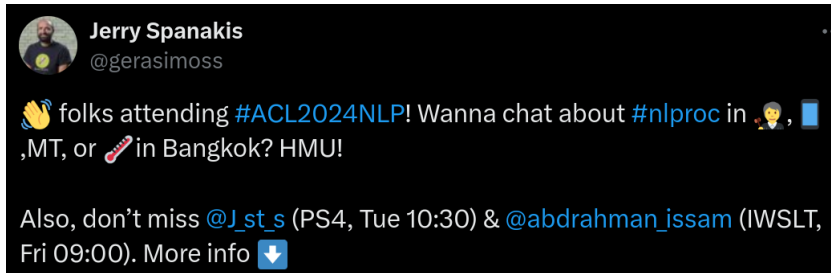
The old men and women gathered by the bank, and they were surprised to see it collapse.

Exercise: find all the ambiguities in this sentence



1. Lexical ambiguity: *bank*
financial institution or edge of a river
2. Structural ambiguity: *The old men and women*
 - Both the men and the women are old.
 - Only the men are old, but the women are not.
3. Semantic ambiguity: *collapse*
financial collapse or physical collapse
4. Anaphoric ambiguity: *they*
 - The old men and women.
 - Just the men, or just the women

Non-standard language



Example tweet including emoji, hashtags, slang (*wanna*), abbreviations (*MT*, *HMU*)

Neologisms

Recent neologisms

- Deepfake
- Prompt engineer
- Situationship
- Enshittification

Neologisms



NYT first said twitter bot 🔗

Need for external knowledge

Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound.

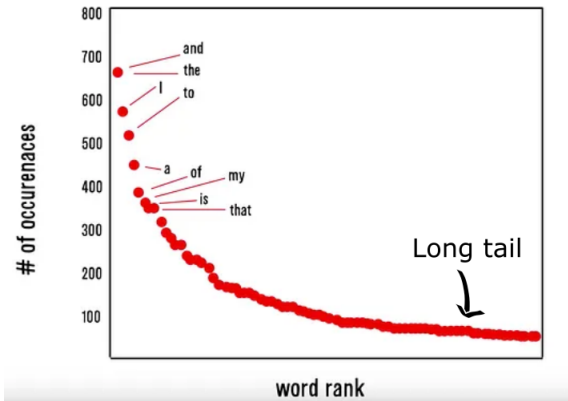
Why was he disappointed?

Minsky (2000)

Types of external knowledge

- Commonsense knowledge of the physical and social world
- Domain knowledge
- Knowledge of the broader context

Law of Zipf



Frequency of words per rank in Romeo and Juliet 🔗

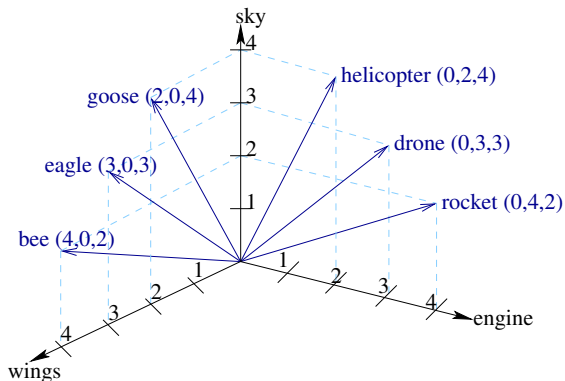
Representing words

Vocabulary: {sky, wings, engine, helicopter, drone, rocket, goose, eagle, bee}

The goose is in the sky.
The eagle in the sky has wings.
The bee in the sky has wings.
The goose with wings in the sky.
The eagle with wings in the sky.
The bee with wings in the sky.
The goose and the eagle with wings are in the sky.
The goose flies in the sky.
The wings of the bee are good wings.
The helicopter, the drone and the rocket are in the sky.
The helicopter in the sky is a useful engine.
The drone in the sky is an engine that takes pictures.
A rocket is an engine.
The helicopter is an engine.
A drone is an engine that takes pictures.
A rocket is a useful engine among sky engines.
A helicopter is flying in the sky.
The sky is full of helicopters and drones.
The engine here is a drone, the one there is a rocket.

Toy corpus

Representing words



Vector space of word co-occurrences with 3 dimensions [link](#)

Representing phrases/sentences

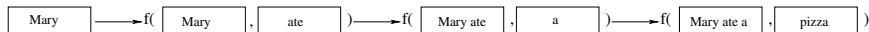
Bag of words

Mary + ate + a + pizza

Mary ate a pizza

Representing phrases/sentences

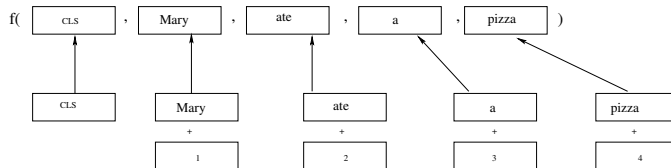
Recurrent Neural Network



Mary ate a pizza

Representing phrases/sentences

Transformer



Mary ate a pizza

Representing phrases/sentences

The reality is not so simple

- Word vectors are learned/updated by the neural network
- $f()$ is complex for RNNs (see lecture 4)
- $f()$ is even more complex for transformers (see lecture 4)
- Neural networks are **deep** and consist of multiple layers.
- The input units of transformers are not words but subwords (see lecture 2)

Fundamentals

1. Introduction
2. Segmentation and tokenization
3. Language modelling
4. Neural language modelling

Levels of linguistic processing

5. POS tagging
6. Morphological analysis
7. Syntactic parsing
8. Semantics
9. Discourse

Applications

10. Neural Machine Translation
11. Question Answering
12. (*Q&A session*)
13. Conversational agents (guest lecture by Thomas Winters)

2. Tokenization

Sample Data:

"This is tokenizing."

Character Level

[T] [h] [i] [s] [i] [s] [t] [o] [k] [e] [n] [i] [z] [i] [n] [g] [.]

Word Level

[This] [is] [tokenizing] [.]

Subword Level

[This] [is] [token] [izing] [.]



- What units to even use?
- How to split text into these units?
- Sentence segmentation

3. Language modelling

Before

$P(\text{I saw a cat on a mat}) =$

- $P(\text{I})$
- $P(\text{saw} \mid \text{I})$
- $P(\text{a} \mid \text{I saw})$
- $P(\text{cat} \mid \text{I saw a})$
- $P(\text{on} \mid \text{I saw a cat})$
- $P(\text{a} \mid \text{I saw a cat on})$
- $P(\text{mat} \mid \text{I saw a cat on a})$

After (3-gram)

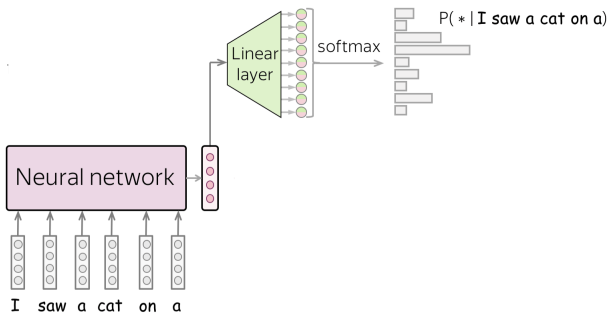
$P(\text{I saw a cat on a mat}) =$

- | | | |
|--|---|-------------------------------------|
| $P(\text{I})$ | → | $P(\text{I})$ |
| • $P(\text{saw} \mid \text{I})$ | → | • $P(\text{saw} \mid \text{I})$ |
| • $P(\text{a} \mid \text{I saw})$ | → | • $P(\text{a} \mid \text{I saw})$ |
| • $P(\text{cat} \mid \text{I saw a})$ | → | • $P(\text{cat} \mid \text{saw a})$ |
| • $P(\text{on} \mid \text{I saw a cat})$ | → | • $P(\text{on} \mid \text{a cat})$ |
| • $P(\text{a} \mid \text{I saw a cat on})$ | → | • $P(\text{a} \mid \text{cat on})$ |
| • $P(\text{mat} \mid \text{I saw a cat on a})$ | → | • $P(\text{mat} \mid \text{on a})$ |
- ignore use



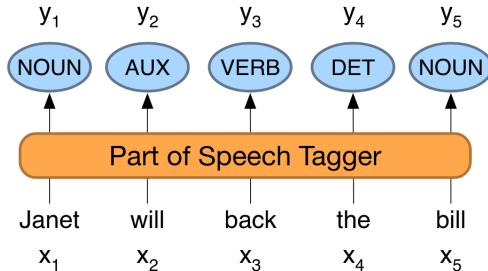
Probabilistic foundations of language modelling

4. Neural Language modelling



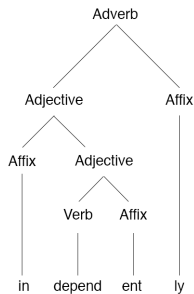
- Language modelling with RNNs
- Language modelling with transformers


5. POS tagging



POS tagging: mapping from input words x_1, x_2, \dots, x_n to output POS tags y_1, y_2, \dots, y_n (Jurafsky and Martin, 2024).

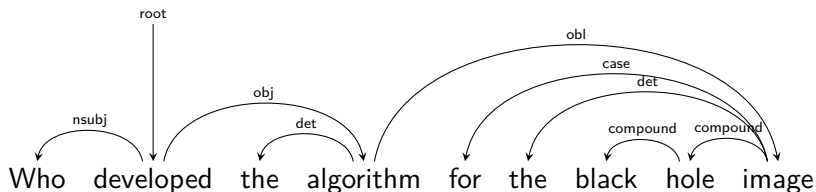
6. Morphological analysis



Morphological analysis of the word *independently* 

Analysing the internal structure of words


7. Syntactic parsing



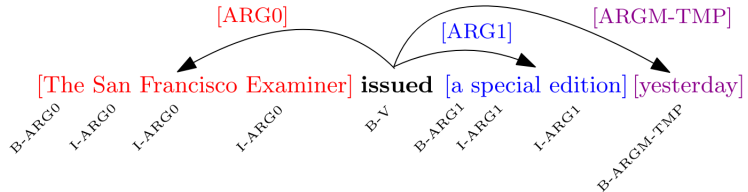
Analysing the structure of sentences

8. Semantics

	sink	
	nail	
	note	
	fan	
	button	

Word sense disambiguation 


8. Semantics



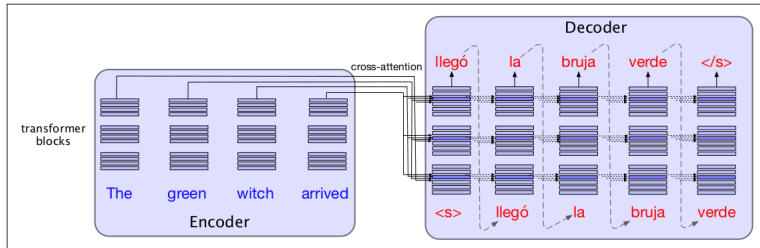
Semantic role labelling: who did what to whom, when (Jindal et al., 2020)

9. Discourse



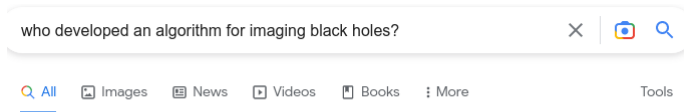
Coreference resolution 

10. Neural Machine Translation



The encoder-decoder transformer architecture for machine translation (Jurafsky and Martin, 2024).

11. Question Answering



About 27.700.000 results (0,36 seconds)

Bouman joined Event Horizon Telescope project in 2013. She led the development of an algorithm for imaging black holes, known as Continuous High-resolution Image Reconstruction using Patch priors (CHIRP).

12. Q&A session

Opportunity to ask any question about any of the lectures or the reading material. We can go through some exercises together.



13. Conversational agents



Not part of the course material **but** fun and informative lecture about LLM-based conversational agents.

Views on teaching

Views

- Traditional lectures are bad - you learn best by activating your brain  .
- Interaction with teacher and peers boosts learning - you should feel **encouraged** to come to class.
- You should not feel **obliged** to come (stay home if you are sick!) I aim to make the course accessible.

Format

- Online: **all content material** (pre-recorded videos of lectures, slides and required readings).
- In-person lectures: interactive and with short exercises. Opportunity to ask for clarifications, get and give feedback.

Classes

- 13 lectures: Tuesdays 16h00 to 18h00
- 6 exercises sessions of 2 hours, 2 groups.
 - Mix of implementation and paper exercises (algorithms application).
 - No additional material but direct preparation for the exam.

Relation to other courses this academic year

It is **complementary** to the *Linguistics and AI* course taught by Tim Van de Cruys (not a requirement).

Linguistics and AI

Stronger focus on the linguistic formalisms and high-level intuitions of models.

NLP

We go deeper in mathematical modelling and see the different algorithms in more detail.

We are collaborating to make the courses as compatible as possible. Redundancy may still happen. We welcome feedback!

Relation to other courses in the past

- Direct replacement of the NLP course formerly taught by Sien Moens.
 - Meant to be more accessible to non-engineering backgrounds.
- ⇒ Thorough revision of the course material.
- Work in progress, feedback welcome!

Course material

Course slides, video recordings and required readings can be downloaded from the [Toledo platform](#).

Evaluation

- Written, open book theoretical part of the exam (50%):
Broad overview questions, comparison of technologies.
- Written, open book exercise part of the exam (50%):
Apply or implement (pseudo-code) algorithms or evaluation metrics and analyse the results.

Teaching team

Course coordinator and lecturer

Miryam de Lhoneux

Dpt of Computer Science

Celestijnlaan 200A, room 4.50

email: miryam.delhoneux@kuleuven.be

Please include [H02B1A] in the subject line

TAs

Kushal Tatariya kushaljayesh.tatariya@kuleuven.be

Wessel Poelman wessel.poelman@kuleuven.be

References

- Ishan Jindal, Ranit Aharonov, Siddhartha Brahma, Huaiyu Zhu, and Yunyao Li. 2020. [Improved semantic role labeling using parameterized neighborhood memory adaptation](#). *ArXiv*, abs/2011.14459.
- Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 20, 2024.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. [Challenges in generalization in open domain question answering](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*.
- Marvin Minsky. 2000. [Commonsense-based interfaces](#). *Commun. ACM*, 43(8):66 – 73.
- Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. [GlobalBench: A benchmark for global progress in natural language processing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- StanfordNLP. 2024. [Cs224n: Natural language processing with deep learning](#).