

Natural Language Processing


Benchmarks/Tasks/Metrics

Neville Walo





🔍 natural | ✕

- 🕒 **natural language processing** Entfern
- 🔍 natural **resources**
- 🔍 natural **disasters**
- 🔍 natural **selection**
- 🔍 natural **history museum**
- 🔍 natural **language processing definition**
-  **Natural Born Killers**
Film (1994)
- 🔍 natural **mojo**
- 🔍 natural **numbers**
- 🔍 natural **gas**

Google Suche Auf gut Glück!

Unangemessene Vervollständigungen melden
[Weitere Informationen](#)



Ungefähr 933'000'000 Ergebnisse (0.62 Sekunden)

Wissenschaftliche Artikel zu **natural language processing**

Natural language processing - Chowdhury - Zitiert von: 944

Natural language processing - Liddy - Zitiert von: 518

Handbook of **natural language processing** - Indurkha - Zitiert von: 887

en.wikipedia.org › wiki › Natural_... ▼ [Diese Seite übersetzen](#)

Natural language processing - Wikipedia

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers ...

[Natural-language understanding](#) · [Natural-language generation](#) · [Sentiment analysis](#)

www.bigdata-insider.de › was-ist-natural-language-proc... ▼

Was ist Natural Language Processing? - BigData-Insider

01.09.2016 — Die Abkürzung **NLP** steht für **Natural Language Processing** und beschreibt Techniken und Methoden zur maschinellen Verarbeitung ...

Ähnliche Fragen

What is meant by natural language processing? ▼

What is natural language processing with example? ▼

What are natural language processing techniques? ▼

What is NLP and its uses? ▼



Natural language processing 🔗

Aus dem Englischen übersetzt - Die Verarbeitung natürlicher Sprache ist ein Teilgebiet der Linguistik, Informatik und künstlichen Intelligenz, das sich mit den Wechselwirkungen zwischen Computern und menschlicher Sprache befasst, ...

[Wikipedia \(Englisch\)](#)

[Ursprüngliche Beschreibung aufrufen](#) ▼

Andere suchten auch nach

[Über 15 weitere ansehen](#)



Maschin...
Lernen



Künstliche
Intelligenz



Computer
Vision



Künstliches
neuronales
Netz



Internet der
Dinge

Text Dokumente

ENGLISCH – ERKANNT DEUTSCH ENGLISCH FRANZÖSISCH ↕ DEUTSCH ENGLISCH FRANZÖSISCH

Natural Language Processing × Verarbeitung natürlicher Sprache ✓

27 / 5000

Feedback geben





Alexa



Siri



Google Now



Cortana

<https://geeksfl.com/blog/best-voice-assistant/>



What is Natural Language Processing?

- **Big picture:** A set of methods and algorithms for making natural languages accessible to computers
- **Analysis** ($NL \rightarrow \mathbb{R}$), e.g., topic classification
- **Generation** ($\mathbb{R} \rightarrow NL$), e.g., chat bots
- **Acquisition** of \mathbb{R} from knowledge and data, e.g., modeling
- \mathbb{R} : some representation interpretable by a machine



Benchmarks

- GLUE (General Language Understanding Evaluation)
- SuperGLUE (Successor of GLUE)

The logo for GLUE, featuring a blue stylized icon of three interconnected nodes to the left of the word "GLUE" in a blue, sans-serif font. The entire logo is centered within a light gray rectangular background.





















GLUE

The logo for SuperGLUE, featuring a red stylized icon of three interconnected nodes to the left of the word "SuperGLUE" in a large, bold, red, sans-serif font.

SuperGLUE







SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

DOWNLOAD ALL DATA



Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b	
+	1	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	93.3/93.8	66.7
+	2	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.2	91.3	95.8/97.6	97.4	88.3/63.0	94.2/93.5	92.7	77.9	95.9	88.8/89.9	66.5
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
+	4	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	92.7/91.9	65.6
+	5	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	87.1/74.4	58.0



Choice of Plausible Alternatives (COPA)

- Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.
- Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.
- Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.



Choice of Plausible Alternatives (COPA)

- Metric: Accuracy

Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b	
+	1	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	URL	90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	93.3/93.8	66.7
+	2	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.2	91.3	95.8/97.6	97.4	88.3/63.0	94.2/93.5	92.7	77.9	95.9	88.8/89.9	66.5
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines	URL	89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
+	4	T5 Team - Google	T5	URL	89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	92.7/91.9	65.6
+	5	Huawei Noah's Ark Lab	NEZHA-Plus	URL	86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	87.1/74.4	58.0



Reading Comprehension over Multiple Sentences (MultiRC)

Paragraph:

Sent 1: Most young mammals, including humans, like to play.

Sent 2: Play is one way they learn the skills that they will need as adults.

Sent 3: Think about how kittens play.

Sent 4: They pounce on toys and chase each other.

Sent 5: This helps them learn how to be better predators.

Sent 6: Big cats also play.

Sent 7: The lion cubs pictured below are playing.

Sent 8: At the same time, they are also practicing their hunting skills.

Sent 9: The dogs are playing tug-of-war with a toy.

Sent 10: What do you think they are learning by playing together this way?

Sent 11: Human children learn by playing as well.

Sent 12: For example, playing games and sports can help them learn to follow rules.

Sent 13: They also learn to work together.

Sent 14: The young child pictured below is playing in the sand.

Sent 15: She is learning about the world through play.

Sent 16: What do you think she might be learning?

Question: What do human children learn by playing games and sports?

- to follow rules
- They learn to follow rules and work together.
- They learn about the world
- Learn to work together
- skills that they will need as adult
- they learn about how to cheat
- how to hunt
- tug-of-war
- only learns to follow rules
- only learns working together
- hunting skills



Reading Comprehension over Multiple Sentences (MultiRC)

- Metric: F1a / EM (Exact Match)

Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b	
+	1	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	↗	90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	93.3/93.8	66.7
+	2	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.2	91.3	95.8/97.6	97.4	88.3/63.0	94.2/93.5	92.7	77.9	95.9	88.8/89.9	66.5
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines	↗	89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
+	4	T5 Team - Google	T5	↗	89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	92.7/91.9	65.6
+	5	Huawei Noah's Ark Lab	NEZHA-Plus	↗	86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	87.1/74.4	58.0



Words in Context (WiC)

Label	Target	Context-1	Context-2
F	bed	There's a lot of trash on the <u>bed</u> of the river	I keep a glass of water next to my <u>bed</u> when I sleep
F	land	The pilot managed to <u>land</u> the airplane safely	The enemy <u>landed</u> several of our aircrafts
F	justify	<u>Justify</u> the margins	The end <u>justifies</u> the means
T	beat	We <u>beat</u> the competition	Agassi <u>beat</u> Becker in the tennis championship
T	air	<u>Air</u> pollution	Open a window and let in some <u>air</u>
T	window	The expanded <u>window</u> will give us time to catch the thieves	You have a two-hour <u>window</u> of clear weather to finish working on the lawn



Words in Context (WiC)

- Metric: Accuracy

Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b	
+	1	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	URL	90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	93.3/93.8	66.7
+	2	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.2	91.3	95.8/97.6	97.4	88.3/63.0	94.2/93.5	92.7	77.9	95.9	88.8/89.9	66.5
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines	URL	89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
+	4	T5 Team - Google	T5	URL	89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	92.7/91.9	65.6
+	5	Huawei Noah's Ark Lab	NEZHA-Plus	URL	86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	87.1/74.4	58.0



Machine Translation

- Task: (NL -> NL)
- Metric: ?

German

Auf dem Tisch ist eine Katze.

English

There is a cat on the table.

On the table is a cat.

There's a cat on the table.

A cat is on a table.

On the table is a kitty cat.

On the table is a bird.

On the desk there is a cat.

Cat on table.

There are cats on the table.

There was a cat on the table.



Machine Translation

- Task: (NL -> NL)
- Metric:
 - BLEU (Bilingual Evaluation Understudy)
 - ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
 - METEOR (Metric for Evaluation of Translation with Explicit ORdering)



N-Gram

- An N-gram is a contiguous sequence of N items from a given sample of text or speech.
- The items can be phonemes, syllables, letters, words or base pairs according to the application.

Name	N	Example
Unigram	1	A
Bigram	2	AB
Trigram	3	UNO
4-gram	4	DATA
N-gram	N	



N-gram Precision

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

$$\text{Unigram Precision} = \frac{m}{w_t} = \frac{7}{7} = 1$$

$$\text{Bigram Precision} = \frac{m}{w_t} = \frac{0}{1} = 0$$

- m : N-grams from the candidate that are found in the reference
- w_t : N-grams in the candidate



BLEU

- Main idea: The closer a machine translation is to a professional human translation, the better it is.
- Scores are calculated over sentences by comparing them with a set of good quality reference translations.
- Scores are then averaged over the whole text to output a final score in $[0,1]$. (1 good translation, 0 bad translation)



BLEU

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

$$\text{Unigram - Precision} = \frac{\sum_w \min(m_w, m_w^{\max})}{w_t} = \frac{2}{7}$$

- m_w : N-grams from the candidate that are found in the reference for N-gram w
- m_w^{\max} : For each N-gram in the candidate translation, the algorithm takes its maximum total count in any of the reference translations.
- w_t : N-grams in the candidate



BLEU

Candidate	the	cat					
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

$$\text{Unigram – Precision} = \frac{\sum_w \min(m_w, m_w^{\max})}{w_t} = \frac{2}{2} = 1$$

$$\text{Bigram – Precision} = \frac{\sum_w \min(m_w, m_w^{\max})}{w_t} = \frac{1}{1} = 1$$

Problems:

- Favors short translations
- Adding more reference translation increases score
- Score \neq Quality

Solutions:

- Punish too short translations
- Consider also recall



ROUGE

- ROUGE-N
- ROUGE-L
- ROUGE-W
- ROUGE-S



ROUGE-N

ROUGE-N (Recall)

$$= \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}_{\text{match}}(gram_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}(gram_n)} \quad (1)$$

BLEU (Precision)

$$\frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{-gram}' \in C'} \text{Count}(n\text{-gram}')}$$



ROUGE-N

Candidate 1	I	always	invariably	perpetually	do
Candidate 2	I	always	do		
Reference 1	I	always	do		
Reference 2	I	invariably	do		
Reference 3	I	perpetually	do		

$$\text{Candidate 1: Unigram Recall} = \frac{9}{9} = 1$$

$$\text{Candidate 2: Unigram Recall} = \frac{7}{9} = 0.77\bar{7}$$



ROUGE-N

Candidate 1	I	always	invariably	perpetually	do
Candidate 2	I	always	do		
Reference 1	I	always	do		
Reference 2	I	invariably	do		
Reference 3	I	perpetually	do		

$$ROUGE - N_{Multi} = \max_{r \in References} ROUGE - N(r, candidate)$$

$$\text{Candidate 1: } ROUGE - 1_{Multi} = \frac{3}{3} = 1$$

$$\text{Candidate 2: } ROUGE - 1_{Multi} = \frac{3}{3} = 1$$



ROUGE

- ROUGE-L: Longest Common Subsequence
- ROUGE-W: Weighted Longest Common Subsequence
- ROUGE-S: Skip-Bigram Co-Occurrence Statistics



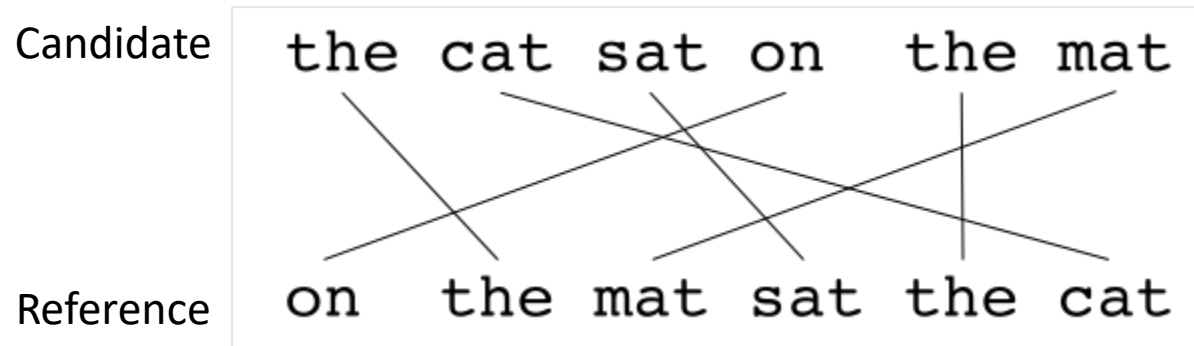
METEOR

- Successor of BLEU
- Based on weighted harmonic mean of unigram precision and recall
- Stemming
 - fishing, fished, fisher = fish
- Synonym Matching
 - buy = purchase



METEOR

- Find maximum alignment with the fewest crosses



$$P = \frac{m}{w_t} = \frac{6}{6} = 1$$

$$R = \frac{m}{w_r} = \frac{6}{6} = 1$$

- m : Number of connections between candidate and reference
- w_r : Number of Unigrams in the reference
- w_t : Number of Unigrams in the candidate



METEOR

- Combine precision and recall

$$F_{mean} = \frac{10PR}{R + 9P}$$

- Calculate penalty p on number of crosses

$$M = F_{mean}(1 - p)$$

- If there are more reference solutions, select highest score
 - Same as ROUGE



Is there a limit?

- GPT-3
 - General language model
 - 175 billion parameters
 - Turing test?



Giving GPT-3 a Turing Test – Common Sense

Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many legs does a frog have?

A: A frog has four legs.

Q: Are there any animals with three legs?

A: No, there are no animals with three legs.

Q: Why don't animals have three legs?

A: Animals don't have three legs because they would fall over.



Giving GPT-3 a Turing Test – Common Sense

Q: How do you sporgle a morgle?

A: You sporgle a morgle by using a sporgle.

Q: How many bonks are in a quoit?

A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?

A: It takes two rainbows to jump from Hawaii to seventeen.

Q: Which colorless green ideas sleep furiously?

A: Ideas that are colorless, green, and sleep furiously are the ideas of a sleep furiously.

Q: Do you understand these questions?

A: I understand these questions.



Giving GPT-3 a Turing Test – Logic

Q: When counting, what number comes before 100?

A: Ninety-nine comes before 100.

Q: When counting, what number comes before 123?

A: One hundred and twenty-two comes before 123.

Q: When counting, what number comes before 1000?

A: Nine hundred and ninety-nine comes before 1000.

Q: When counting, what number comes before 10,000?

A: Nine thousand and ninety-nine comes before 10,000.

Q: When counting, what number comes before a million?

A: Nine hundred thousand and ninety-nine comes before a million.

Q: Write one line of Ruby code to reverse an array.

A: `ary.reverse`

Q: Write one line of Ruby code to count how many X's are in an array.

A: `ary.count { |x| x == 'X' }`

Q: Reverse the following array: [1, 3, 5, 6, 10, 4, 2, 77]

A: [10, 6, 4, 2, 77, 3, 5, 1]

Q: How many X's are in this array: [1, 3, 5, 6, 10, 4, 2, 77]

A: 2



Giving GPT-3 a Turing Test – Logic

Q: If I have a marble and a paper clip in a box, put a pencil in the box, and remove the marble, what is left?

A: A paper clip.

Q: If I have two shoes in a box, put a pencil in the box, and remove one shoe, what is left?

A: A shoe.

Q: If I put a pencil in a box, then put another pencil in the box, what is in the box?

A: Two pencils.

Q: Then if I take out a pencil and put in a frog, what is in the box?

A: A frog.



Giving GPT-3 a Turing Test

- Some of the problems can be solved by using a different prompt
 - E.g. If the question is nonsense say «get real»



Conclusion

- NLP is everywhere
- Benchmarks: GLUE and SuperGLUE
 - State-of-the-art performance
- Metrics for machine translation
 - BLEU
 - ROUGE
 - METEOR
- Limitations of GPT-3

