

自然言語処理における アーカイブの現状と課題

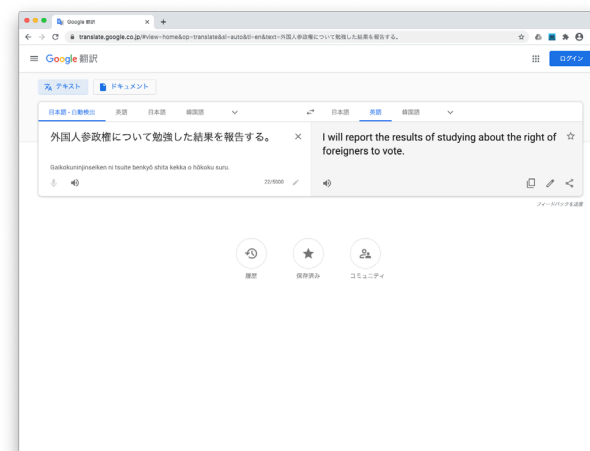
河原 大輔

早稲田大学 理工学術院

自然言語処理とは

(自然)言語をコンピュータで処理することに関する学問、技術

- 基礎解析
 - 形態素解析、構文解析など
- 応用システム
 - 情報検索、機械翻訳、対話システムなど

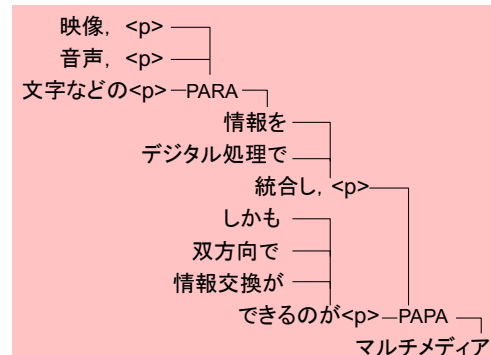


映像、音声、文字などの情報をデジタル処理で統合し、しかも双方向で情報交換ができるのがマルチメディア。

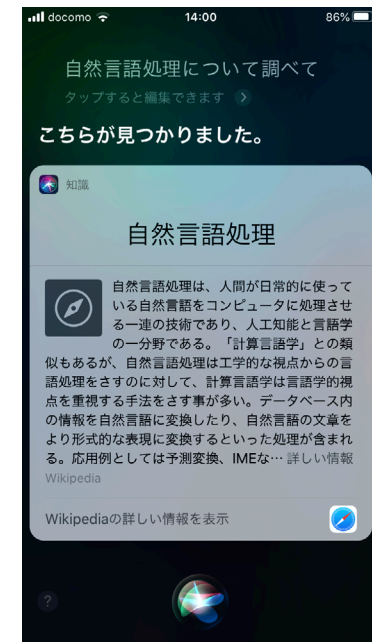


形態素解析

映像 えいぞう 映像 名詞 普通名詞 **
... 特殊 読点 **
音声 おんせい 音声 名詞 普通名詞 **
... 特殊 読点 **
文字 もじ 文字 名詞 普通名詞 **
など など 助詞 副助詞 **
の の 助詞 接続助詞 **
情報 じょうほう 情報 名詞 普通名詞 **
を を 助詞 格助詞 **
デジタル でじたる デジタル 名詞 普通名詞 **
処理 しょり 処理 名詞 サ変名詞 **
で で 助詞 格助詞 **
統合 とうごう 統合 名詞 サ変名詞 **
ししする 動詞 * サ変動詞 基本連用形
... 特殊 読点 **
しかも しかも しかも 接続詞 **
...



構文解析



自然言語処理におけるアーカイブ

≡ コーパス (corpus)

コーパス

- もともとはある主題に関する文書や、ある作者の文書を集めたもの
- 現在ではもう少し広く、文書、または音声データを集めたもの、またそこに種々の情報を付与したもの

生コーパス: 言語情報を人手で付与していないもの

- 新聞記事コーパス
 - 新聞社各社が年単位で販売
 - 1年あたり100万文程度
- Webコーパス
 - Common Crawl (<https://commoncrawl.org/>)
 - 毎月、2週間程度でクロールしたWebページ集合
 - 30億ページ(各国語)、300TB
 - cf. Internet Archive (<https://archive.org/>)
- Wikipedia本文
 - 英語 (<https://dumps.wikimedia.org/enwiki/latest/>)
 - 620万記事 (1億文)
 - 日本語 (<https://dumps.wikimedia.org/jawiki/latest/>)
 - 120万記事 (2000万文)

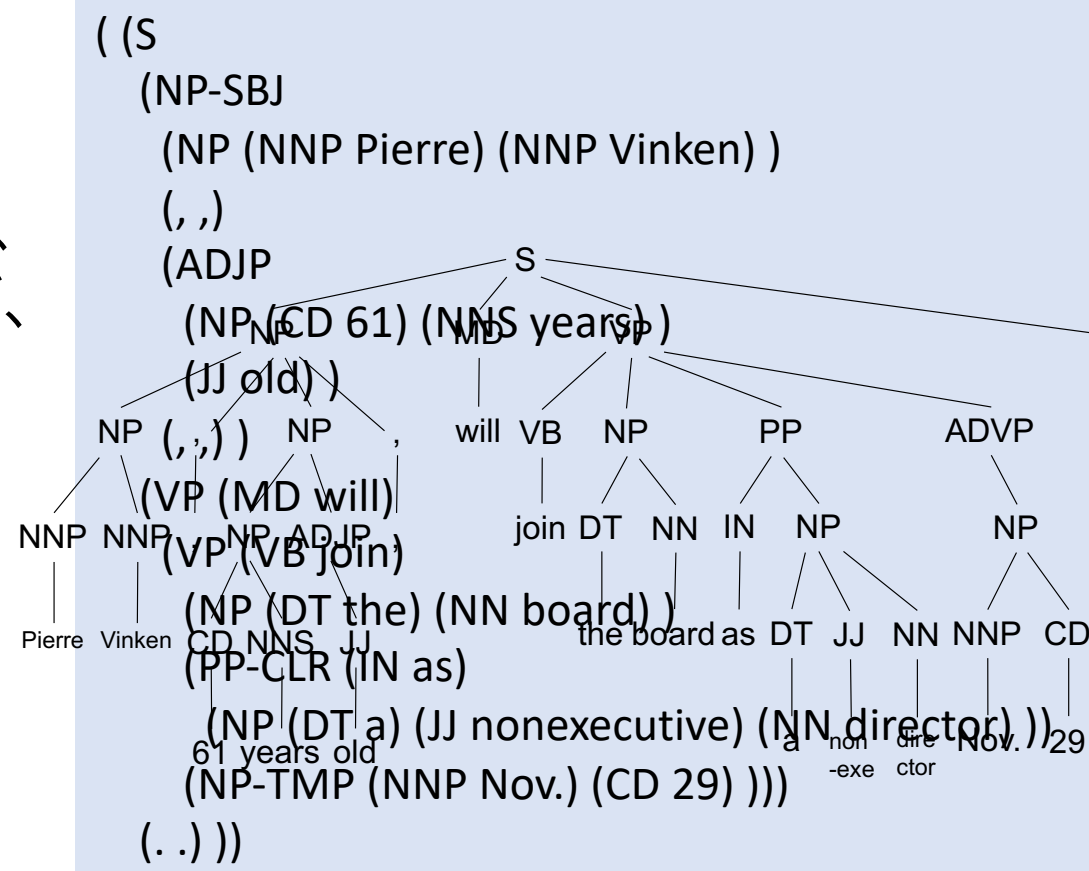
注釈付与コーパス: 言語情報を人手で付与したもの

→ 機械学習の教師データ、評価データとして用いられる

- 構文情報
 - Penn Treebank [Marcus+ 1993]
 - 京都大学テキストコーパス [Kurohashi & Nagao 1998]
- 意味情報
 - 自然言語推論
 - SNLI [Bowman+ 2015]
 - 文章読解
 - SQuAD [Rajpurkar+ 2016]
 - 意味理解に関するベンチマークデータセット
 - GLUE [Wang+ 2018]

Penn Treebank [Marcus+ 1993]

- Wall Street Journalの1989年記事約100万語に対して品詞・構文情報を付与
- 括弧の入れ子構造によって構造を表現し、NP(名詞句)、VP(動詞句)などの構文タグ、NN(名詞)、VB(動詞)などの品詞タグを付与
- Linguistic Data Consortium (LDC) から配布
 - <https://catalog.ldc.upenn.edu/LDC99T42>



言語資源の配布機関

- アメリカ
 - Linguistic Data Consortium (LDC)
- ヨーロッパ
 - European Language Resources Association (ELRA)
- 日本
 - 言語資源協会 (GSK)
 - 高度言語情報融合フォーラム (ALAGIN)
 - 情報学研究データリポジトリ (IDR)

京都大学テキストコーパス [Kurohashi & Nagao 1998]

- 毎日新聞の1995年記事約4万文 (約100万語)に単語区切り、品詞・読み・活用、文節区切り、係り受け情報を付与
- GitHubで配布 (付与情報のみ)
 - <https://github.com/ku-nlp/KyotoCorpus>
 - 毎日新聞CD-ROMを購入する必要あり

# S-ID	村山D
* 24C	富市D
+ 1D	首相は
村山	
+ 2D	年頭にD
富市	
+ 36C	あたり
首相	
はは	首相D
* 2D	官邸で
+ 4D	内閣D
年頭	
にに	記者D
* 6D	会と
+ 11C	二十八日D
あた	
* 6D	会見し、
+ 6D	社会党の
首相	
+ 11C	新D
官邸	
でで	民主D
...	連合D

図形 8

注釈付与コーパス: 言語情報を人手で付与したもの

→ 機械学習の教師データ、評価データとして用いられる

- 構文情報
 - Penn Treebank [Marcus+ 1993]
 - 京都大学テキストコーパス [Kurohashi & Nagao 1998]
- 意味情報
 - 自然言語推論
 - SNLI [Bowman+ 2015]
 - 文章読解
 - SQuAD [Rajpurkar+ 2016]
 - 意味理解に関するベンチマークデータセット
 - GLUE [Wang+ 2018]

The Stanford Natural Language Inference (SNLI) Corpus [Bowman+ 2015]

- 57万文ペアからなる自然言語推論のコーパス
- 2文間の関係として含意、矛盾、中立の3つのラベルをクラウドソーシングによって付与

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

- Web配布: <https://nlp.stanford.edu/projects/snli/>

The Stanford Question Answering Dataset (SQuAD) [Rajpurkar+ 2016]

- 10万問からなる文章読解問題セット
- Wikipediaの一段落に対する質問と答えをクラウドソーシングで作成

Text: Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, ...

Q: What is the second most abundant element?

A: helium

- Web配布:
<https://rajpurkar.github.io/SQuAD-explorer/>

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language https://arxiv.org/abs/1810.04805	87.433	93.160
2 May 14, 2019	ATB (single model) Anonymous	86.940	92.641
3 Feb 14, 2019	Knowledge-enhanced BERT (single model) Anonymous	85.944	92.425
4 Feb 28, 2019	ST_bl single model	85.430	91.976
4 Sep 26, 2018	nlnet (ensemble) Microsoft Research Asia	85.954	91.677
5 Feb 16, 2019	BERT+Sparse-Transformer single model	85.125	91.623
5 Sep 09, 2018	nlnet (ensemble) Microsoft Research Asia	85.356	91.202
5 Oct 05, 2018	BERT (single model) Google AI Language https://arxiv.org/abs/1810.04805	85.083	91.835
5 Mar 14, 2019	BISAN (single model) Seoul National University & Hyundai Motors	85.314	91.756

GLUE (General Language Understanding Evaluation)

[Wang+ 2018]

一文		タスク名	説明
文ペア	一文	SST-2	映画レビューのポジネガ判定
		CoLA	文が言語的に正しいかどうか
	文ペア	MRPC	2文が同じ意味かどうか
		STS-B	2文の類似度 (1から5)
		QQP	2質問文が同じ意味かどうか
		MNLI	2文が含意/矛盾/中立か
		QNLI	SQuADを変形し、質問と文が与えられ、文が答えを含むかどうか
		RTE	含意関係かどうか
		WNLI	Winograd Scheme Challengeから作られている。含意関係かどうか

GLUE (Generalized Language Understanding Evaluation)

[Wang+ 2018]

タスク名

SST-2	映画レビューのポジネガ判定
CoLA	文が言語的に正しいかどうか
MRPC	2文が同じ意味かどうか
STS-B	2文の類似度 (1から5)

[positive]

it's a charming and often affecting journey.

[negative]

a sometimes tedious film.

[同じ]

The top rate will go to 4.45 percent for all residents with taxable incomes above \$500,000.
For residents with incomes above \$500,000, the income-tax rate will increase to 4.45 percent.

[異なる]

While dioxin levels in the environment were up last year, they have dropped by 75 percent since the 1970s, said Caswell.
The Institute said dioxin levels in the environment have fallen by as much as 76 percent since the 1970s.

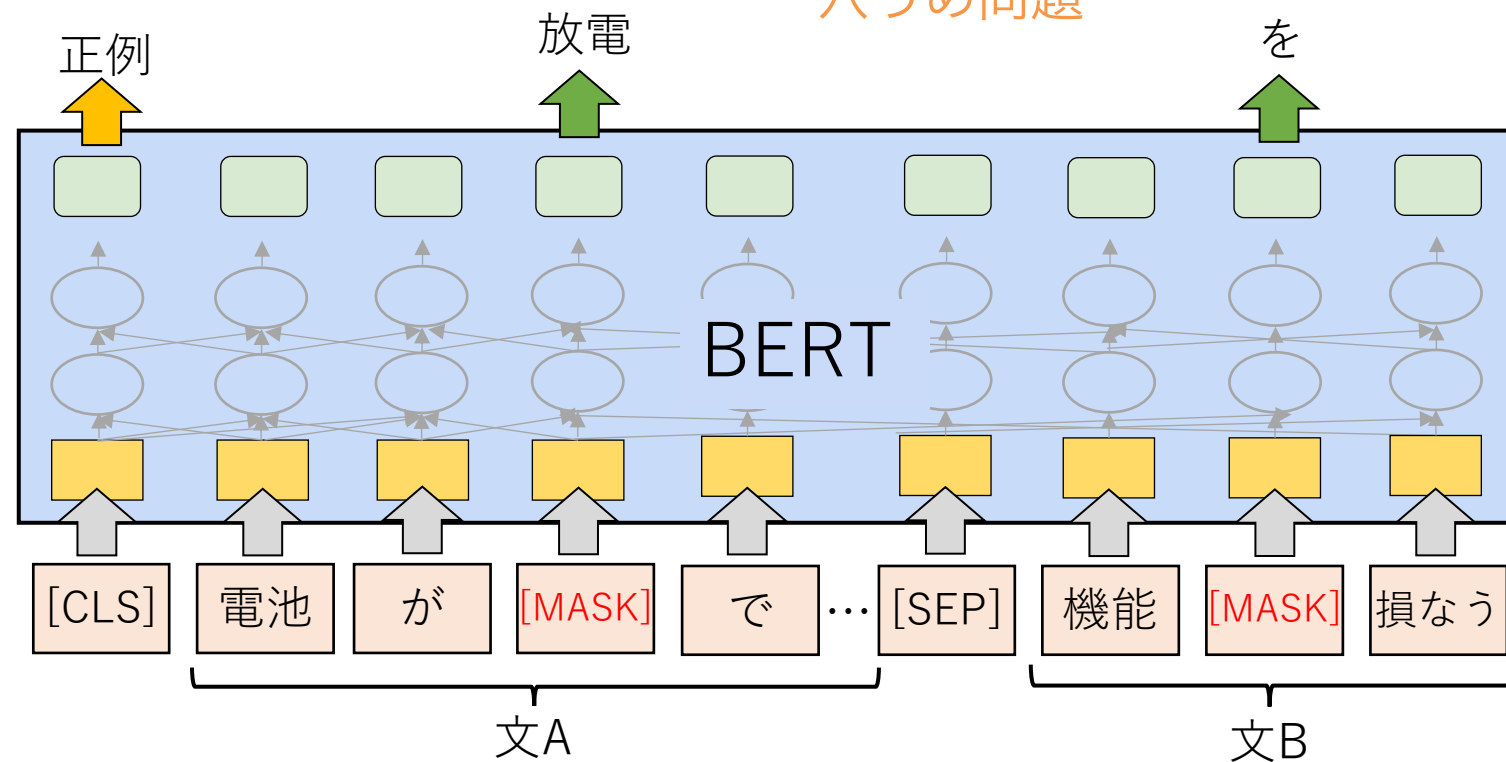
か

コーパスの利用例: BERT [Devlin+ 2019]

1. Pre-training

文A, Bが隣接文
かどうかを識別

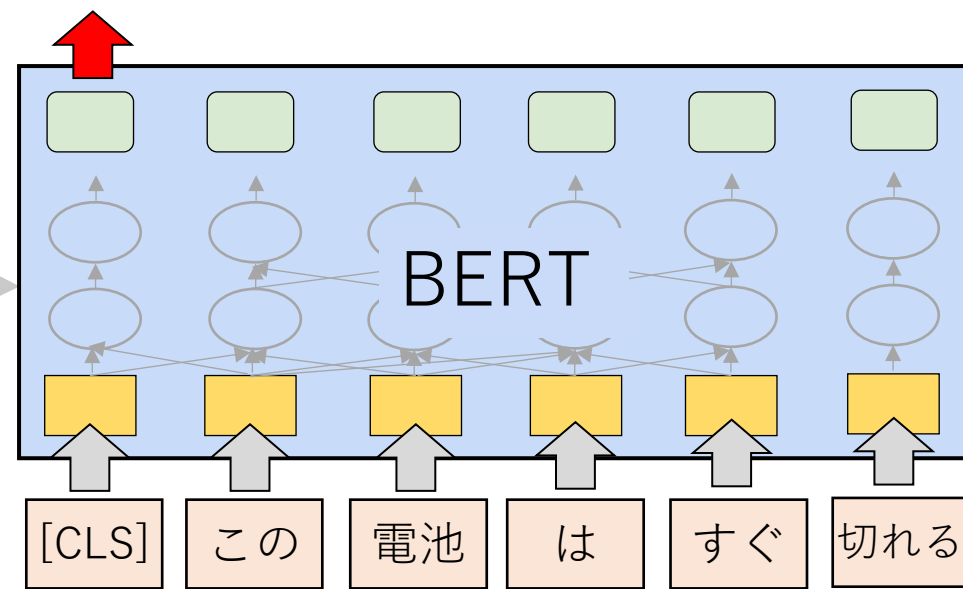
“穴うめ問題”



2. Fine-tuning

タスク: ポジネガ判定

negative



GLUE Leaderboard

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	T5 Team - Google	T5	🔗	89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	53.1
2	ALBERT-Team Google Language	ALBERT (Ensemble)	🔗	89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
+	3 王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)	🔗	89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7	90.2	99.2	87.3	89.7	47.8
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	🔗	88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
5	Facebook AI	RoBERTa	🔗	88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	90.8	90.2	98.9	88.2	89.0	48.7
6	XLNet Team	XLNet-Large (ensemble)	🔗	88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	90.2	89.8	98.6	86.3	90.4	47.5
+	7 Microsoft D365 AI & MSR AI	MT-DNN-ensemble	🔗	88.4	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
8	GLUE Human Baselines	GLUE Human Baselines	🔗	87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
9	Stanford Hazy Research	Snorkel MeTaL	🔗	83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
10	XLM Systems	XLM (English only)	🔗	83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
11	Zhuosheng Zhang	SemBERT	🔗	82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
12	Danqi Chen	SpanBERT (single-task training)	🔗	82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
13	Kevin Clark	BERT + BAM	🔗	82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
14	Nitish Shirish Keskar	Span-Extractive BERT on STILTs	🔗	82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
15	Jason Phang	BERT on STILTs	🔗	82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
16	廖亿	RGLM-Base (Huawei Noah's Ark Lab)	🔗	80.5	56.9	94.2	90.7/87.7	89.7/89.1	72.2/89.4	86.1	85.4	92.1	78.5	65.1	40.0
+	17 Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden	🔗	80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
18	Neil Houlsby	BERT + Single-task Adapters	🔗	80.2	59.2	94.3	88.7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2
19	Zhuohan Li	Macaron Net-base	🔗	79.7	57.6	94.0	88.4/84.4	87.5/86.3	70.8/89.0	85.4	84.5	91.6	70.5	65.1	38.7
20	蘇大約	SesameBERT-Base	🔗	78.6	52.7	94.2	88.9/84.8	86.5/85.5	70.8/88.8	83.7	83.6	91.0	67.6	65.1	35.8
+	21 MobileBERT Team	MobileBERT	🔗	78.5	51.1	92.6	88.8/84.5	86.2/84.8	70.5/88.3	84.3	83.4	91.6	70.4	65.1	34.3
22	Linyuan Gong	StackingBERT-Base	🔗	78.5	51.1	92.6	88.8/84.5	86.2/84.8	70.5/88.3	84.3	83.4	91.6	70.4	65.1	34.3
23	Huawei Noah's Ark Lab	TinyBERT (4-layers; 7.5x smaller and 9.4x faster than BERT-base)	🔗	78.5	51.1	92.6	88.8/84.5	86.2/84.8	70.5/88.3	84.3	83.4	91.6	70.4	65.1	34.3
24	shijing si	bert+pos6	🔗	78.5	51.1	92.6	88.8/84.5	86.2/84.8	70.5/88.3	84.3	83.4	91.6	70.4	65.1	34.3
25	GLUE Baselines	BiLSTM+ELMo+Attn	🔗	70.0	33.6	90.4	84.4/78.0	74.2/72.3	63.1/84.3	74.1	74.5	79.8	58.9	65.1	21.7

人間のスコア: 87.1

BERT: 80.5

ベースライン: 70.0

注釈付与コーパス: 言語情報を人手で付与したもの

→ 機械学習の教師データ、評価データとして用いられる

- 構文情報
 - Penn Treebank [Marcus+ 1993]
 - 京都大学テキストコーパス [Kurohashi & Nagao 1998]
- 意味情報
 - 自然言語推論
 - SNLI [Bowman+ 2015]
 - 文章読解
 - SQuAD [Rajpurkar+ 2016]
 - 意味理解に関するベンチマークデータセット
 - GLUE [Wang+ 2018]

日本語コーパスに関する課題

- 特にアジアにおいて、ポータルやプラットフォームが整備されていない
- 日本語Webコーパスのデータは基本的に公開されていない
- 注釈付与コーパスの構築、共有が英語、中国語圏と比べて遅れている