Proc. ACL 2023 Volume 1: Long Papers, pages 6825–6845

Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization

Rongxin Zhu Jianzhong Qi Jey Han Lau

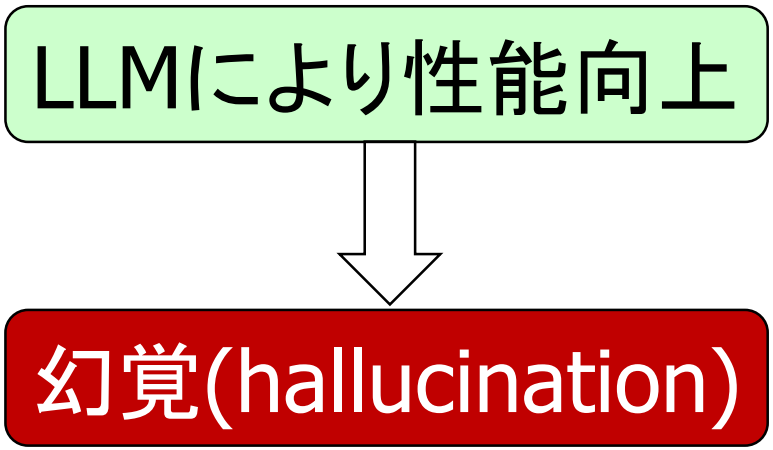
School of Computing and Information Systems
The University of Melbourne

紹介者：小川 泰弘（名市大）

自動要約

入力された文書を簡潔にまとめる

- 抽出型要約(extractive summarization)
  - 重要文だけ抜き出す
- 抽象型要約(abstractive summarization)
  - 複数の文をまとめ、新たな文を生成



事実誤認 (Factual Error)

- 自動要約における課題
- その分類や検出がホットトピック
  - 今回13本

紹介の動機

Question Answering-2

補正予算によるコロナ対策

〔1〕医療提供体制及び経済活動と都民生活を万全の体制で守り抜くべき。〔2〕感染状況を踏まえ保健所への派遣職員の増員を。〔3〕自宅療養の支援体制を拡大すべき。〔4〕中小企業等を対象とした多様な支援策を継続し必要な財源措置を講じよ。〔5〕大規模イベントの検査ルールを策定し、感染症の拡大防止と経済活動の両立を図るべき

知事 充実、[redacted]全の対[redacted]策定や[redacted]

総務局長 〔2〕都職員120名程度常時派遣している。規模の拡大含めしっかりと対応図る。

担当局長 〔3〕都のLINEアプリの健康管理や食料品配送等、希望する区市へ導入進める。

産業労働局長 〔4〕事業実施期間の再度の延長や、感染状況に応じた支援措置を検討する。

Newsletter

check

Answer Verification

check

答弁1

▼知事

答弁に先立ちまして、一言弔意を申し上げます。

名誉都民である有馬朗人さんが逝去されました。ここに謹んで哀悼の意を表し、心よりご冥福をお祈りいたします

Answers

What is QA Lab-PoliInfo-4 ?

Diet or local assembly

Person asking a question

Person answering the question

Assembly minutes

○議長（石川良一君）　これより質問に入ります。
百十五番小山くにひこ君。
（百十五番小山くにひこ君登壇）

○百十五番（小山くにひこ君）　令和二年第四回定例会に当たり、都民ファーストの会東京都議団を代表し、小池知事及び教育長、関係局長に質問いたします。
初めに、過日、名誉都民である小柴昌俊さん、有馬朗人さんがご逝去されました。ここに謹んで哀悼の意を表し、心よりご冥福を[redacted]
質問[redacted]お亡くなりになられ[redacted]申し上げます。ま[redacted]一日も早いご回復を祈念申し上げます。
国内外において、新型コロナの第三波というべき状況が到来しています。東京でも多くの新規陽性者が発生し、重症者数の推移も予断を許さない状況です。医療崩壊を起こさせないため、新型コロナ対応に当たってくださっている医療従事者、医療機関への支援を一層強化しながら、これまでの知見を踏まえた、めり張りのついた対策を進め、感染拡大の防止をまずはしっかりと行いながら、社会経済活動との両立を図っていく必要があります。

Transcript

Stance Classification-2

	議案15号	議案16号	議案17号	議案18号	議案19号
山口広文					
深谷美登					
岩城荘平					
度島剛一					
大西勝彦					
上西正雄					
浅田茂彦					
鈴置英昭					
窪地洋	賛成	反対	反対	賛成	賛成

Approve or disapprove

Minutes-to-Budget Linking

歳入の精査等（２，７４２億円）

- **都税等** ▲１，９５５億円
  新型コロナウイルス感染症の影響に伴う企業収益の悪化等により減収となります。
  - **都税収入等の状況**

区分	令和２年度最終補正後	令和２年度当初予算	増▲減
都税	5兆２，５２５億円	5兆４，４４６億円	▲１，９２１億円
地方譲与税	４９５億円	５２９億円	▲３４億円
合計	5兆３，０２０億円	5兆４，９７５億円	▲１，９５５億円

- **減収補填債の発行** １，０００億円
- **繰越金** ９９３億円
- **国庫支出金（東京2020大会追加経費負担分）** ７１０億円

Budget table

本論文の目的と貢献

- 詳細で文レベルの事実誤認の注釈付き対話コーパスの作成 (DiaSumFact)
- SOTAモデルでの事実誤認の検出性能の調査と困難さの提示
- 文書要約における事実誤認検出の新手法の提案(BertMulti, EnDeRanker)
- 様々な事実誤認検出手法の長所と短所の分析

DiaSumFact (誤りタグ付き対話コーパス)

データソースと要約生成モデル

- SAMSum (Gliwa et al.,2019)
  - 日常会話と正解要約
  - BART, ConDigSum, GPT-3 , PEGASUS, S-BART
- QMSum (Zhong et al., 2021)
  - 会議議事録の質問と回答。回答が要約
  - BART, DialogLM, PEGASUS

- 各60対話
- 12人による注釈

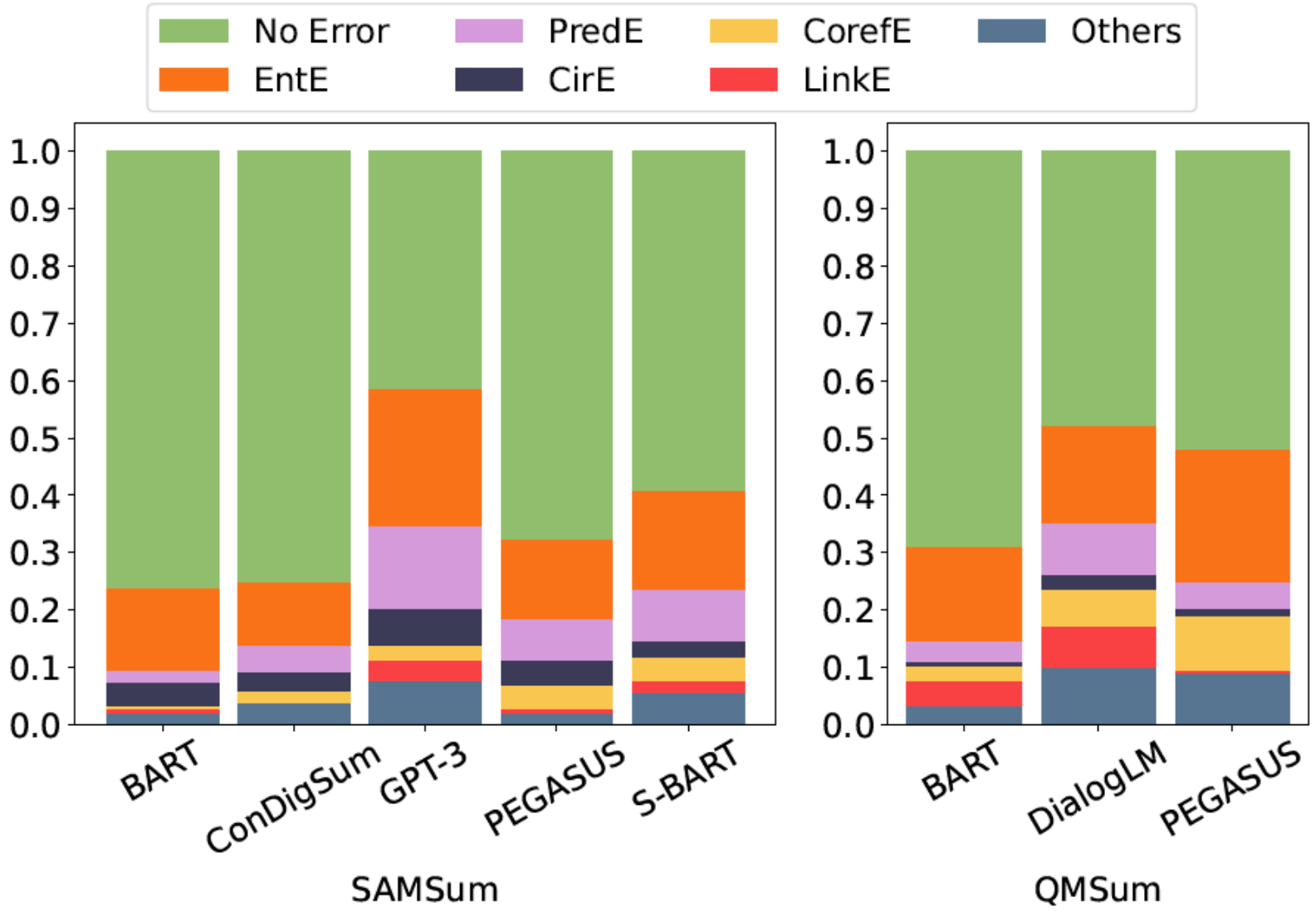
事実誤認の分類

マルチラベル・2次元

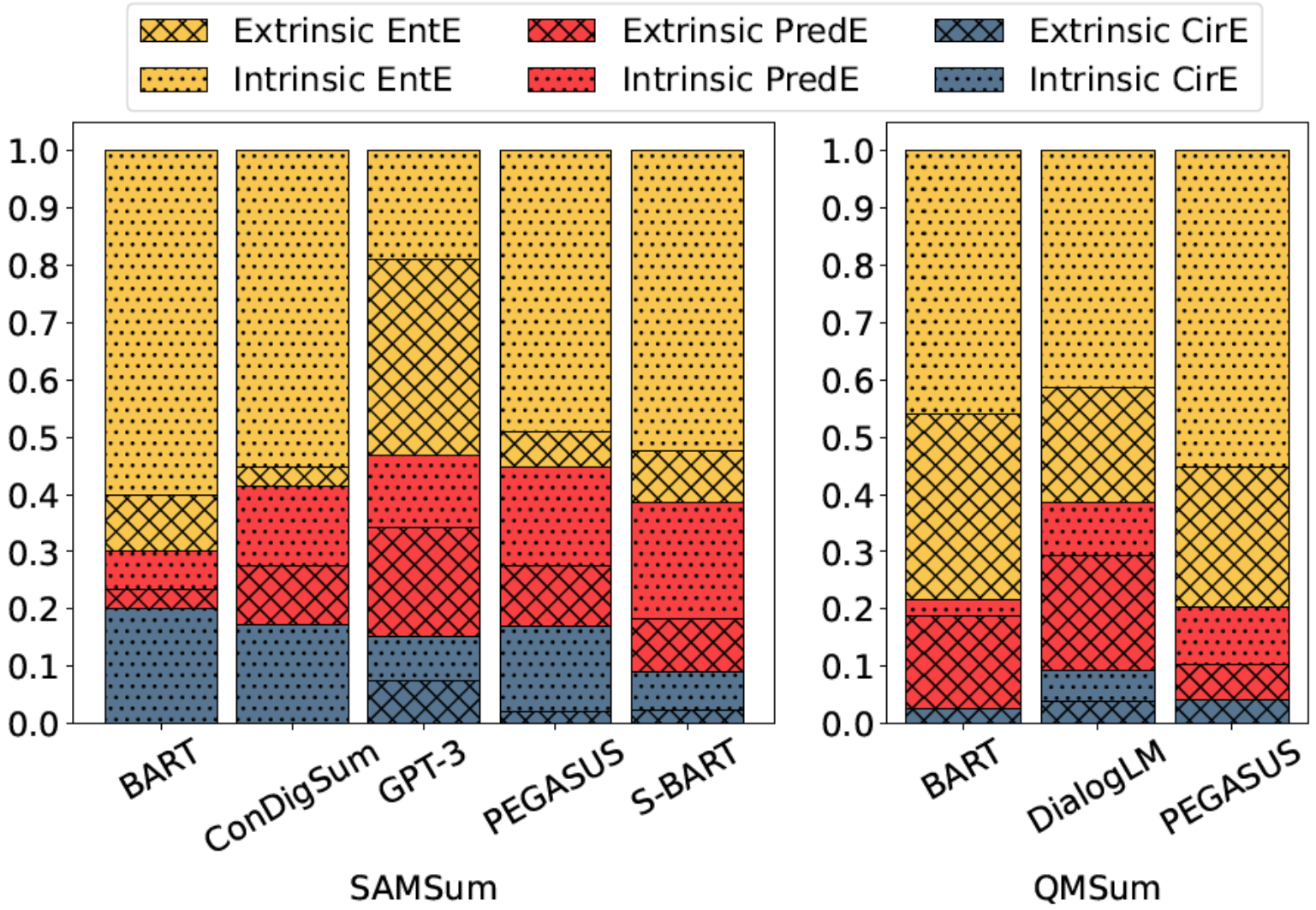
	対話		
	Lucas: Where r u? I'm waiting at the airport. Vanessa: There was a foul-up with the flight. I'm trying to get another ticket. Lucas: OMG. How come? Vanessa: No bloody idea. All of the flights are booked cos students are returning from holidays. Lucas: I've called the airport and they said there's a flight to New York at 9:45 p.m. Vanessa: Great, I'll book it now.		
	誤り型	詳細	例文	In/EX
semantic frame	EntE	主語・目的語の誤り	*Vanessa* is waiting at the airport.	In
semantic frame	PredT	述語の誤り	Lucas *has emailed* the airport and got some information about the flight to New York.	Ex
semantic frame	CirE	修飾語の誤り	Lucas is waiting *at the train station*.	Ex
discourse	CorefE	代名詞・参照の誤り	Vanessa is trying to get another ticket *for themselves*.	N/A
discourse	LinkE	言明間の関係の誤り	Vanessa will book the flight to New York at 9:45 pm because students are returning from holidays.	N/A
	Other	その他		N/A

Intrinsic/Extrinsic: 誤り部分が原文書にあるかないか

誤りの分布1


No Error
PredE
CorefE
Others
EntE
CirE
LinkE
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
BART
ConDigSum
GPT-3
PEGASUS
S-BART
SAMSum
BART
DialogLM
PEGASUS
QMSum

誤りの分布2


Extrinsic EntE
Intrinsic EntE
Extrinsic PredE
Intrinsic PredE
Extrinsic CirE
Intrinsic CirE
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
BART
ConDigSum
GPT-3
PEGASUS
S-BART
SAMSum
BART
DialogLM
PEGASUS
QMSum

事実誤認の検出

- QAFactEval (Fabbri et al., 2022)
  - QAベース
  - 要約中の名詞句や固有表現を隠して問題作成。その問題に原文書を利用して回答。元の表現と答えの類似度で判定
- DAE (Goyal and Durrett, 2020)
  - 原文書から要約中の依存関係が含意できるかで判定
- BertMulti
  - Bertを使った弱教師あり学習によるマルチクラス分類器

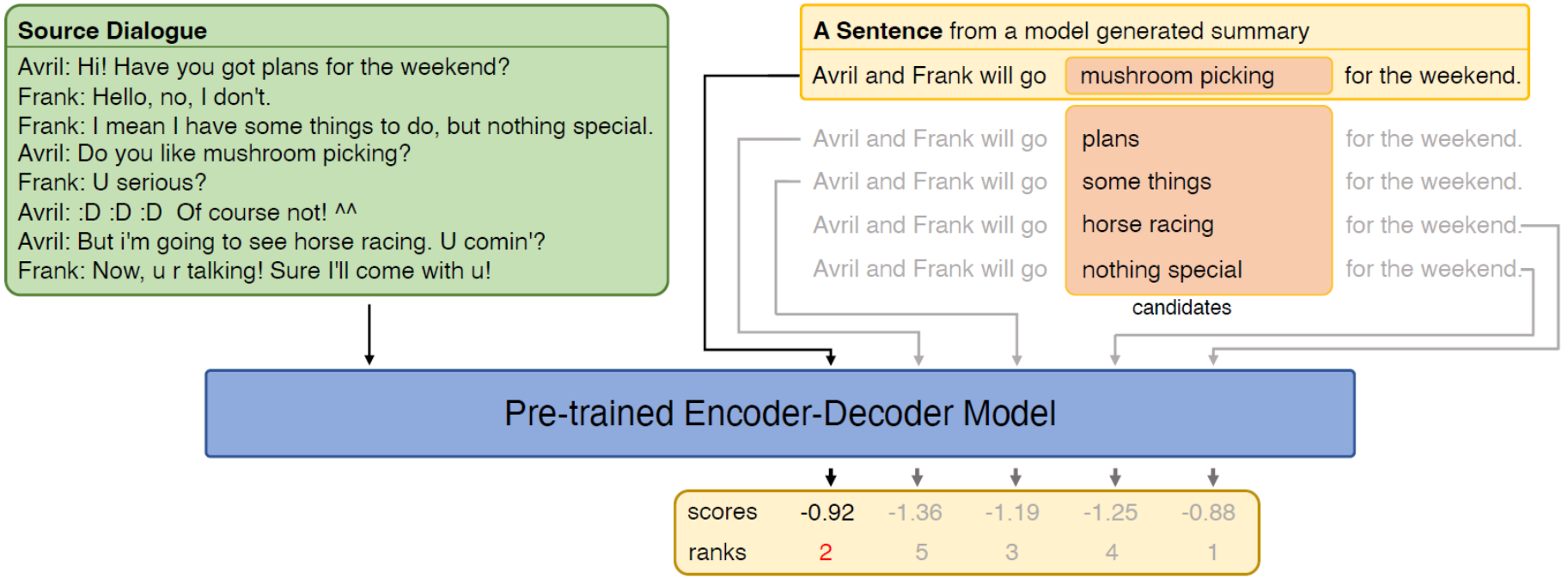
事実誤認の検出 提案手法

- EnDeRanker
  - ➢教師無しモデル
  1. SOI (spans of interest)を求める
     - ✧候補は名詞句・固有表現・動詞
  2. SOIを原文書中の同等の語で置換
  3. それぞれを Encoder-Decoderモデルを通してスコアを計算

$$\text{score}_c = \frac{1}{n}\sum_{i=1}^{n} \log p(w_i | w_{<i}, D)$$

  4. スコアが最大でなければ事実誤認と判定

事実誤認の検出 提案手法　例



mushroom picking is not factually consistent, because its rank is 2, larger than T=1 (assuming T=1. T is a tunable hyper-parameter).

事実誤認の検出結果

Model	NoE	EntE	CirE	PredE	CorefE	Others	Micro Avg	Macro Avg
Adapted state-of-the-art models								
QAFACTEVAL	$0.68_{0.04}$	$\underline{0.45}_{0.03}$	$\underline{0.23}_{0.11}$	$0.00_{0.00}$	$0.11_{0.06}$	$0.00_{0.00}$	$0.51_{0.03}$	$0.25_{0.02}$
DAE	$0.77_{0.02}$	$0.32_{0.05}$	$0.03_{0.06}$	$0.00_{0.00}$	$0.00_{0.00}$	$\underline{0.34}_{0.11}$	$0.59_{0.02}$	$0.24_{0.02}$
Weakly Supervised multi-class classifier								
BERTMULTI	$0.72_{0.00}$	$0.20_{0.00}$	$0.08_{0.00}$	$0.09_{0.00}$	$\underline{0.29}_{0.00}$	$0.08_{0.00}$	$0.54_{0.00}$	$0.24_{0.00}$
ENDERANKER (ours)								
BART-LARGE-CNN	$0.67_{0.06}$	$0.34_{0.07}$	$0.04_{0.06}$	$0.15_{0.04}$	$0.12_{0.10}$	$0.00_{0.00}$	$0.47_{0.07}$	$0.22_{0.01}$
BART-LARGE-SAMSUM	$0.67_{0.06}$	$0.35_{0.08}$	$0.03_{0.04}$	$0.21_{0.06}$	$0.21_{0.13}$	$0.00_{0.00}$	$0.47_{0.05}$	$0.24_{0.02}$
PEGASUS-CNN	$0.71_{0.03}$	$0.37_{0.08}$	$0.04_{0.05}$	$0.18_{0.05}$	$0.14_{0.09}$	$0.00_{0.00}$	$0.52_{0.04}$	$0.24_{0.01}$
PEGASUS-SAMSUM	$0.67_{0.04}$	$0.37_{0.09}$	$0.06_{0.07}$	$0.19_{0.06}$	$0.16_{0.11}$	$0.01_{0.02}$	$0.46_{0.05}$	$0.24_{0.01}$
T5-LARGE-CNN	$0.68_{0.04}$	$0.35_{0.09}$	$0.03_{0.04}$	$0.15_{0.04}$	$0.06_{0.03}$	$0.01_{0.02}$	$0.47_{0.05}$	$0.21_{0.02}$
T5-LARGE-SAMSUM	$0.70_{0.08}$	$0.35_{0.10}$	$0.04_{0.05}$	$\underline{0.22}_{0.08}$	$0.14_{0.03}$	$0.00_{0.00}$	$0.51_{0.09}$	$0.24_{0.03}$
Ensemble learning (**including our ENDERANKER model**)								
FREQVOTING	$0.79_{0.03}$	$0.40_{0.05}$	$0.05_{0.11}$	$0.10_{0.08}$	$0.12_{0.10}$	$0.01_{0.02}$	$\underline{0.62}_{0.03}$	$0.24_{0.03}$
LOGISTIC	$\underline{0.80}_{0.03}$	$0.44_{0.05}$	$0.20_{0.13}$	$0.00_{0.00}$	$0.11_{0.10}$	$0.03_{0.03}$	$0.61_{0.03}$	$\underline{0.26}_{0.04}$

その他のデータセット

Dataset	#Mod	#Summ	#Sen	Domain	Annotation Typology
FactCC (Kryscinski et al., 2020)	10	/	1,434	news	binary (consistent, inconsistent)
QAGS (Wang et al., 2020)	2	474	/	news	binary (consistent, inconsistent)
SummEval (Fabbri et al., 2021)	44	12,800	/	news	5-point Likert scale
Polytope (Huang et al., 2020)	10	1,500	/	news	multi-class
Cao'22 (Cao et al., 2022)	1	800	/	news	multi-class
Maynez'20 (Maynez et al., 2020)	5	500	/	news	binary (intrinsic, extrinsic)
Frank (Pagnoni et al., 2021)	8	2,250	4,942	news	multi-class
Goyal'21 (Goyal and Durrett, 2021)	3	50	/	news	multi-dimensional, multi-class
CLIFF (Cao and Wang, 2021)	2	600	/	news	multi-class
ConFIT (Tang et al., 2022b)	4	76	/	dialogue	multi-class
DialSummEval (Gao and Wan, 2022)	13	4,200		dialogue	5-point Likert Scale
DIASUMFACT (ours)	6	475	1,340	dialogue	multi-dimensional, multi-class

まとめ

- 要約における事実誤認について様々な研究
- 本論文では、
  - 事実誤認をマルチラベル・2次元に分類
  - 誤りタグ付き対話コーパスの作成
  - アンサンブルを含めて、すべての種類の事実誤認に有効な方法はない