

# ACL参加報告

水本 智也

フューチャー株式会社

2019年9月28日  
第15回 テキストアナリティクス・シンポジウム





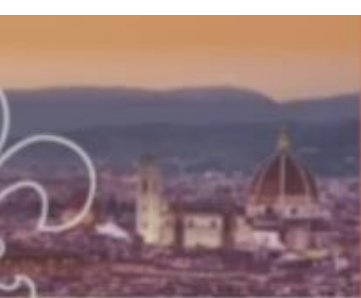
# ACL2019会議概要



# ACL2019

- 自然言語処理の難関国際会議
- 57回目の開催
- 開催場所：イタリア、フィレンツェ
- 開催時期：7/28-8/2
  - 7/28: チュートリアル
    - 9件
  - 7/29-7/31: 本会議
    - 700件近くの発表
    - 6並列+ポスター
  - 8/1-8/2: ワークショップ
    - 19ワークショップ





57<sup>th</sup>

ACL 2019

ANNUAL MEETING

of the Association for Computational Linguistics

Florence (Italy)

July 28<sup>th</sup> - August 2<sup>nd</sup>

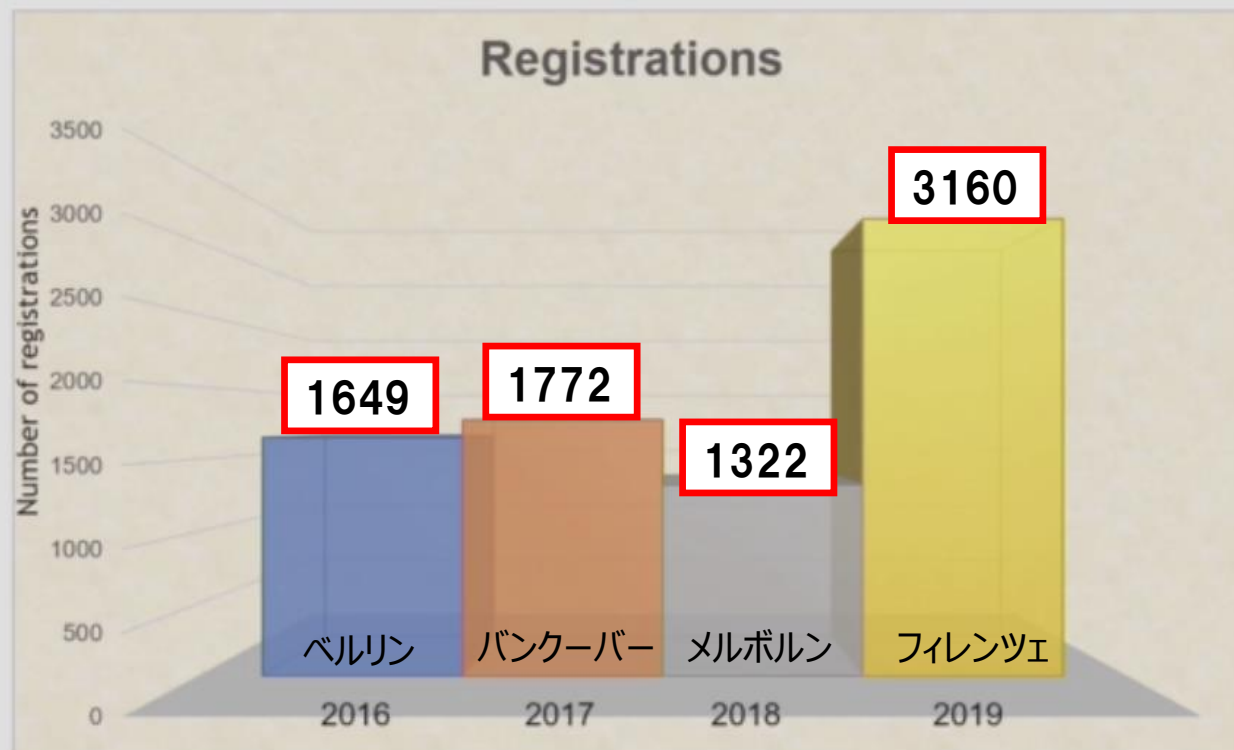
Fortezza da Basso

ACL size went **XXL**

\* ACL2019 Opening slideより

# 参加者数の推移

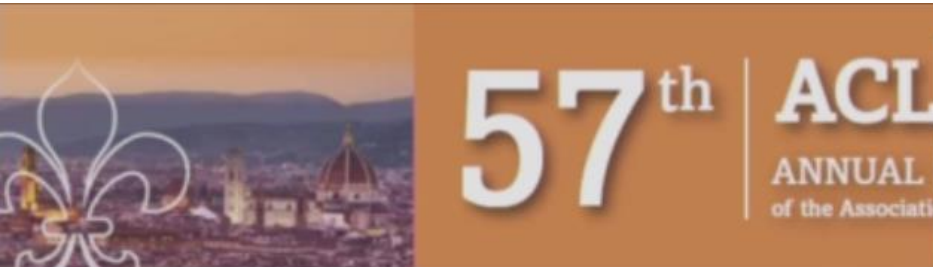
- 2018年参加者の**2倍以上**



\* ACL2019 Opening slideより

# 投稿数の推移

- 2018年投稿数の約2倍



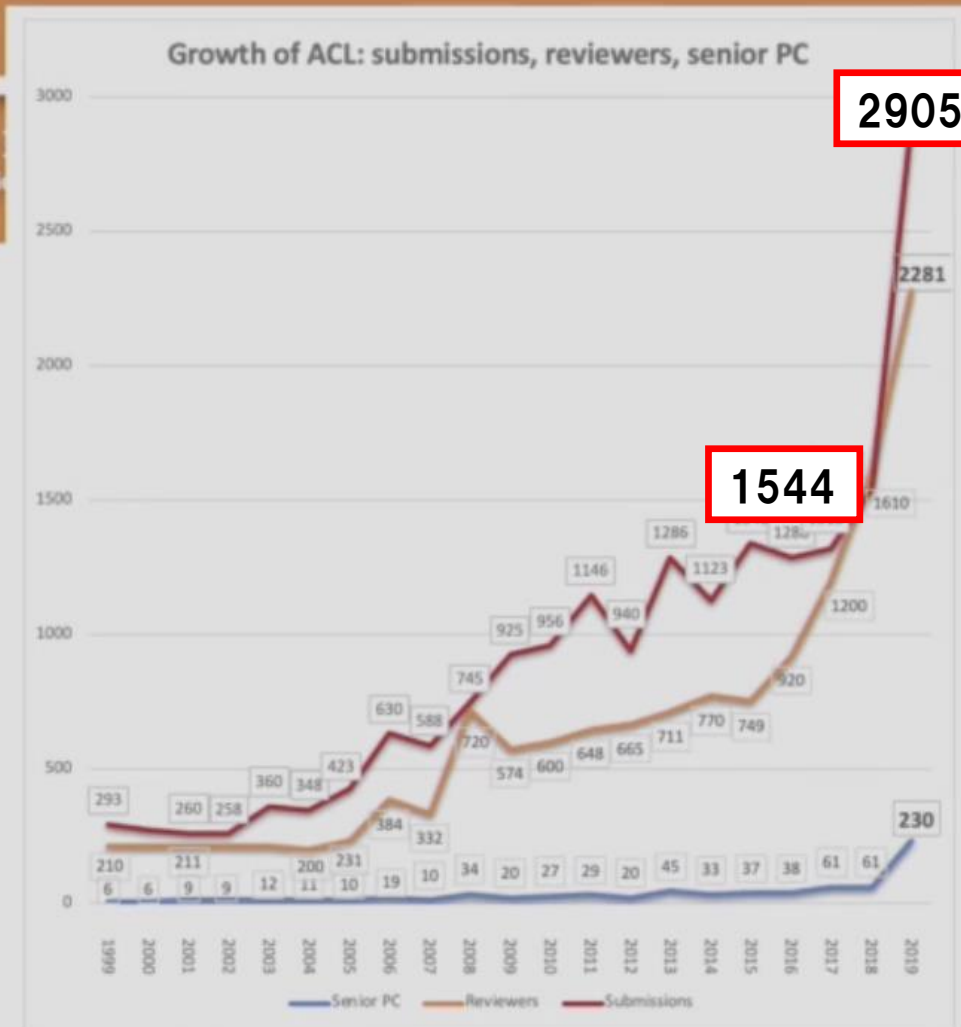
## Submissions

A 75% increase over ACL 2018!

An all-time record for ACL-related Conferences

Submissions from 74 countries/regions  
incl. a few from Antarctica 😊

2,694 valid submissions (1,609 long and  
1,085 short papers) underwent review



\* ACL2019 Opening slideより

# 採択率

- 採択率は例年通り、Accept本数が激増



## Acceptance rates

Conference		Submissions	Accepts	Accept rate (%)
ACL 2019	All	2905	660	22.7
	Long	1737	447	25.7
	Short	1168	213	18.2
ACL 2018	All	1544	384	24.9
	Long	1018	258	25.3
	Short	526	126	24.0
ACL 2017	All	1297	302	23.3
	Long	737	195	26.5
	Short	560	107	19.1

\* ACL2019 Opening slideより

# ACLトレンド

# 2018

# 2019

# 2017

A word cloud visualization of NLP-related terms. The words are arranged in various sizes and colors (green, blue, purple, orange). Four red rectangular boxes highlight specific groups of words:

- Top right box:** generation, document, inference, text, effective
- Middle left box:** unsupervised, domain, evaluation, machine
- Bottom center box:** summarization, translation, simple, experts, reviews, error
- Bottom left box:** abstractive, embeddings, disentangled



# ベストペーパー

- 会議より前にベストペーパー候補が発表（HPで公開）
  - Long papers 17本
  - Short papers 11本
  - Demo papers 4本
- 本会議最終日のクロージングにて発表



## Best papers

32 nominations - 28 main session papers + 4 demo papers:

<http://www.acl2019.org/EN/nominations-for-acl-2019-best-paper-awards.xhtml>

8 awards:



# Outstanding papers

- Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts テキストマイニング
- A Simple Theoretical Model of Importance for Summarization 要約
- Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems 対話
- We need to talk about standard splits 分析系
- Zero-shot Word Sense Disambiguation using Sense Definition Embeddings 語義曖昧性解消

# Best demo paper

- **OpenKiwi: An Open Source Framework for Quality Estimation**
  - *Fabio Kepler, Jonay Trenous, Marcos Treviso, Miguel Vera and André F. T. Martins*
    - Unbabel, Instituto de Telecomunicacoes
  - 機械翻訳のQuality Estimation (QE)タスクのためのフレームワークを提供
    - QEタスク: 正解データなしで翻訳の質を評価
  - 先行研究の4つの手法を実装し提供
  - 新しいデータで上の4つの手法のモデルを学習することも可能
  - WMT2019のShared Taskのベースラインとしても使用



# Best short paper

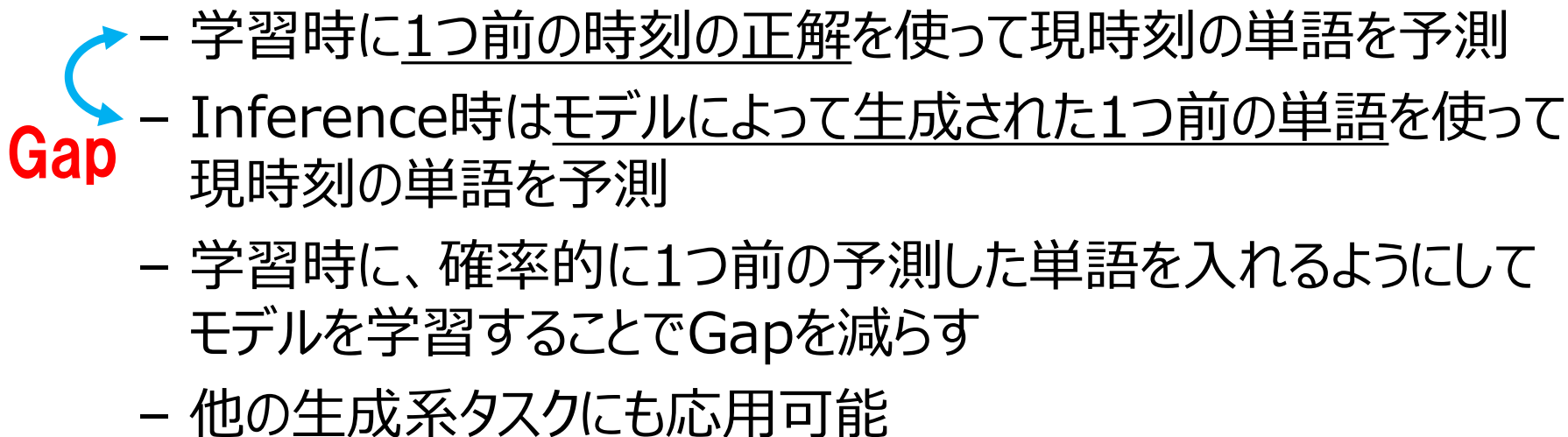
- **Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment**
  - *Nanjiang Jiang and Marie-Catherine de Marneffe*
    - The Ohio State University
  - Speaker commitment: 話者の発言が、ある事象に対してどの程度現実のものなのか、不確かなのかの具合
    - 例: 事象「フィレンツェは観光客でいっぱい」どっちが現実的か**
      - フィレンツェは観光客でいっぱいだと**知っていますか?**
      - フィレンツェは観光客でいっぱいだと**思いますか?**
  - 先行研究の2つのモデルを同一データセットにおいて比較し分析
    - ルールベースモデル v.s. 深層学習モデル
    - ルールベースの方が良かった

# Best long paper

- **Bridging the Gap between Training and Inference for Neural Machine Translation**

- *Wen Zhang, Yang Feng, Fandong Meng, Di You and Qun Liu*

- University of Chinese Academy of Sciences, Tencent Inc, Worcester Polytechnic Institute, Huawei

- 
- 学習時に1つ前の時刻の正解を使って現時刻の単語を予測
  - Inference時はモデルによって生成された1つ前の単語を使って現時刻の単語を予測
  - 学習時に、確率的に1つ前の予測した単語を入れるようにしてモデルを学習することでGapを減らす
  - 他の生成系タスクにも応用可能

# ホットキーワード

- **Bias**

- 人種、性別など
- 1セッション＋ワークショップ

- **Health**

- 対話から症状を予測したり、良いカウンセラはどういうものか予測
- 1セッション＋関連ワークショップ

- **Evaluation**

- 生成系のタスクが主流になり、どう評価するかにフォーカス
- Evaluationがつく発表15件

- **Interpretability**

- 解釈性
- 7件＋ワークショップ



# 論文紹介



# 紹介する論文

- **Health**

- Extracting Symptoms and their Status from Clinical Conversations
  - Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran

- **Evaluation**

- HighRES: Highlight-based Reference-less Evaluation of Summarization
  - Hardy, Shashi Narayan and Andreas Vlachos

# EXTRACTING SYMPTOMS AND THEIR STATUS FROM CLINICAL CONVERSATIONS

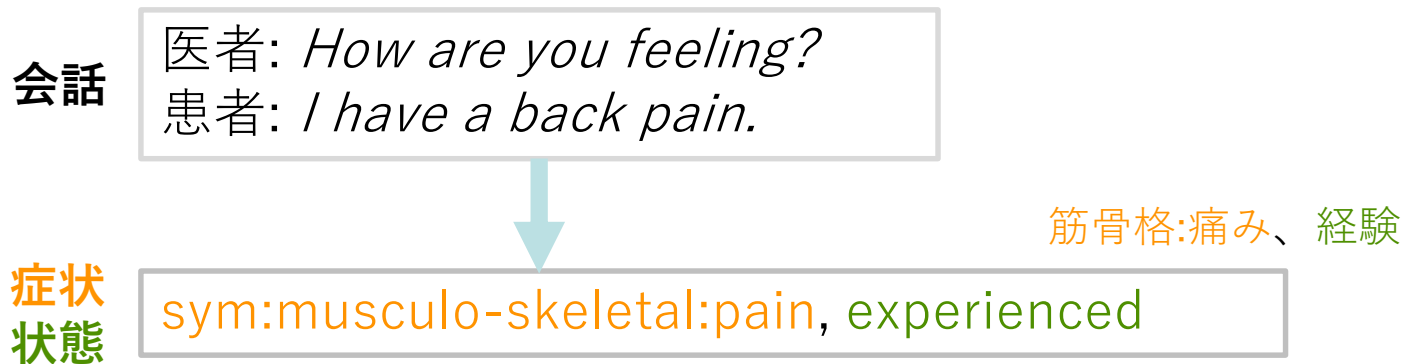
Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen,  
and Izhak Shafran

Google Inc.



# 概要

- **タスク:** 医者と患者の会話からSymptom(症状)とStatus(状態)の予測



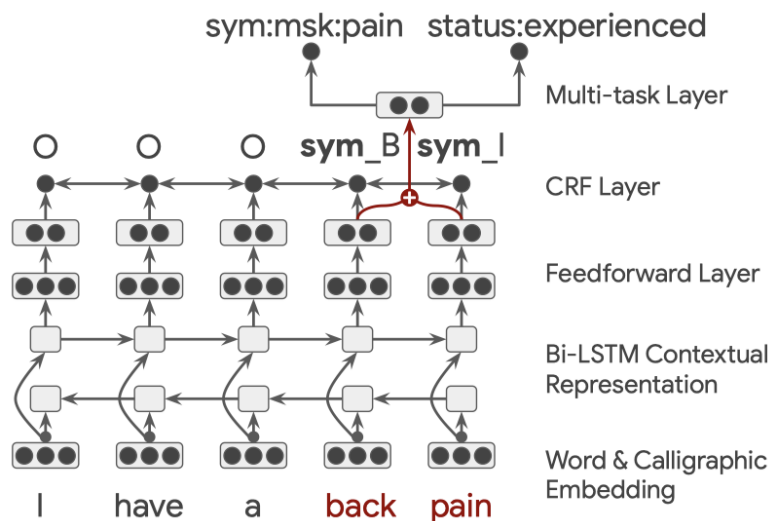
- **背景:** 11時間の勤務時間のうち4.5時間を診療文書作成に使う（通常勤務終了後、さらに1.4時間）
  - この時間を少しでも短くして患者との時間を増やしたい

# データセット

- 元データ: 90,000件の非特定化された書き起こし
  - 1件あたり医者と患者の10分ほどの音声会話
  - 一部は看護師、介護士、配偶者などの音声含む
- 症状と状態のアノテーションスキーマ
  - 医学書士、医者、NLP専門家によって定義
  - 186の症状でそれぞれが体の部位(14部位)と紐づく
  - 3つの状態(experienced, not experienced, other)
- 2950件に対して18人のプロの医学書士がアノテート
  - Inter-labeler agreement (kappa): 0.4
    - 医者と患者の症状の話し方が曖昧もしくはインフォーマル
    - アノテータが関連した似たラベルを付与
    - アノテータの付けたラベルのスパンの不一致

# 2つのモデルを提案

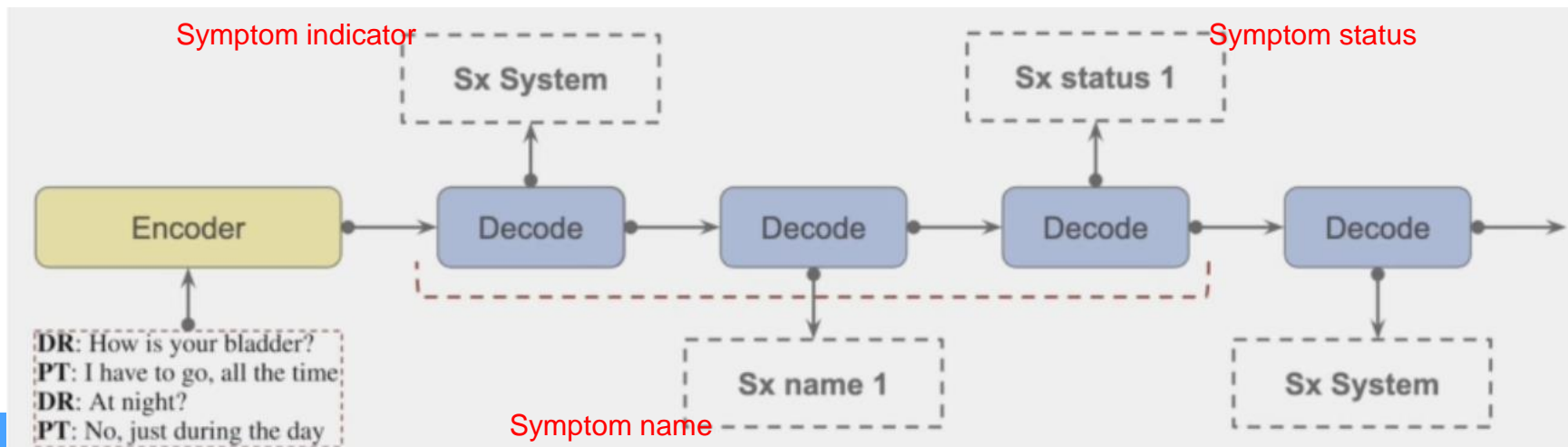
- Span-Attribute Tagging Model (SA-T)



← Symptomの中身とStatusを予測

← Symptomにあたる箇所を同定

- Sequence-to-sequence Model (Seq2Seq)



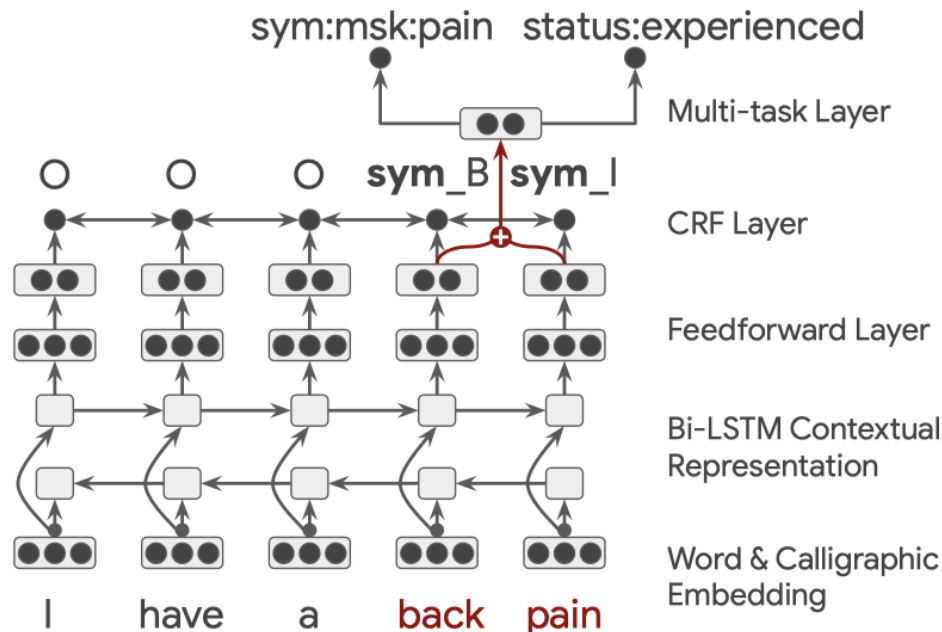
# SA-T Model

- スパンを当てるタスク: CRF (with BIO tags)

I	0
have	0
a	0
back	B-sym:musculo-skeletal:pain:experienced
pain	I-sym:musculo-skeletal:pain:experienced
.	0

症状が186種類、状態が3種類、  
BIで2つ、0が1つ  
 $(186 * 3 * 2) + 1 = 1117$ ラベル

- ラベルの数が多すぎて現実的ではない
  - 二段階に分けて症状と状態を推定



1. CRF-layerで範囲を同定
2. 1で同定された範囲の症状と状態を推定



# Seq2Seq Model

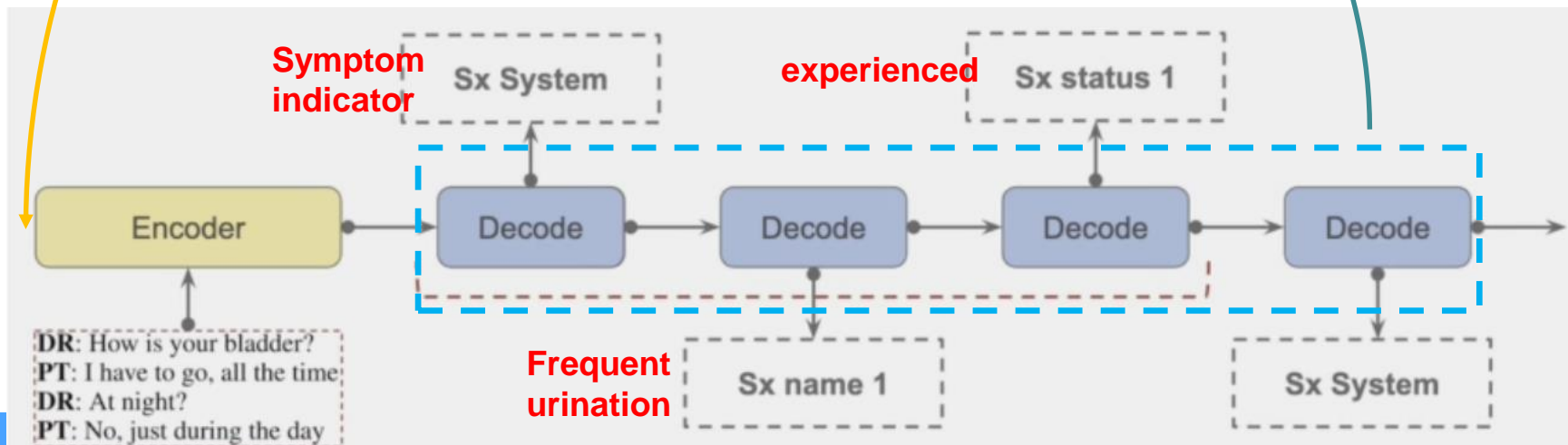
- 症状が明示的に出ない、会話を通して出現

Transcript	Symptoms + Status
<b>DR:</b> Any issues with your eyes? <b>PT:</b> Well sort of <b>DR:</b> Is your vision ok? <b>PT:</b> Yeah, but the right one hurts	<b>Eye pain:</b> experienced <b>Vision loss:</b> not experienced
<b>DR:</b> How is your bladder? <b>PT:</b> I have to go, all the time <b>DR:</b> At night? <b>PT:</b> No, just during the day	<b>Frequent urination:</b> experienced <b>Nocturia:</b> not experienced

入力

出力

Transcriptを入力してSymptoms+Statusを出力するSeq2Seq



# 実験結果

Weighted F1= 複数回出る症状に重みを置いた評価指標

Model	Sx	Sx + Status
<i>Unweighted F1(Precision, Recall)</i>		
Baseline	<b>0.68</b> (0.73, 0.63)	<b>0.50</b> (0.54, 0.47)
SA-T	<b>0.71</b> (0.73, 0.69)	<b>0.58</b> (0.58, 0.58)
Seq2Seq	<b>0.70</b> (0.66, 0.75)	<b>0.55</b> (0.49, 0.62)
<i>Weighted F1(Precision, Recall)</i>		
Baseline	<b>0.73</b> (0.78, 0.69)	<b>0.57</b> (0.61, 0.53)
SA-T	<b>0.77</b> (0.80, 0.74)	<b>0.65</b> (0.66, 0.63)
Seq2Seq	<b>0.79</b> (0.77, 0.80)	<b>0.64</b> (0.61, 0.68)

## 提案モデルの比較

1. SA-TとSeq2Seqが同程度の性能（著者からすると驚き）
  - SA-Tはまず位置を要特定
2. SA-TはPrecisionが高く、Seq2SeqはRecallが高い
3. ベースラインへの言及なし

## 人同士の結果との比較

1. 人でも難しい
2. 人は症状がわかると状態も当てられる
  - 状態の推定性能向上は課題

Model	Sx	Sx + Status
<i>Unweighted F1(Precision, Recall)</i>		
Human	<b>0.84</b> (0.86, 0.82)	<b>0.78</b> (0.80, 0.76)
SA-T	<b>0.71</b> (0.73, 0.69)	<b>0.58</b> (0.58, 0.57)
Seq2Seq	<b>0.70</b> (0.66, 0.75)	<b>0.55</b> (0.49, 0.62)
<i>Weighted F1(Precision, Recall)</i>		
Human	<b>0.86</b> (0.88, 0.85)	<b>0.81</b> (0.82, 0.79)
SA-T	<b>0.77</b> (0.80, 0.74)	<b>0.65</b> (0.66, 0.63)
Seq2Seq	<b>0.79</b> (0.77, 0.80)	<b>0.64</b> (0.61, 0.68)

差が小さい

差が大きい

# HIGHRES: HIGHLIGHT-BASED REFERENCE-LESS EVALUATION OF SUMMARIZATION

Hardy<sup>1</sup>, Shashi Narayan<sup>2</sup> and Andreas Vlachos<sup>1,3</sup>

<sup>1</sup>University of Sheffield, <sup>2</sup>Google Research, <sup>3</sup>University of Cambridge

**Reference = 正解の要約**

# 概要

- 要約の正解文を使わない新しい評価方法の提案
- 要約の正解は、1つのソースに対して1つのみが多い
  - 生成タスクにおいて正解となるものは複数あり得る
- 要約の人手評価は大変
  - 人が評価しても一貫的な評価は難しい
  - 評価する人が変われば結果も変わる
- 元のソースに対して重要なところをマーキング
- その重要単語を元に自動評価
  - 正解はいらないけど、ソースに人手でつけるコスト。。。
  - 一度つければ使い回しは可能

# HIGHRESの3つのコンポーネント

- ハイライトアノテーション
  - ソースドキュメント中の重要な単語、フレーズをハイライト
- ハイライトベース内容評価
  - 全ての重要な情報が要約に入っているか (Recall)
  - 重要な情報のみが要約に入っているか (Precision)
- 明瞭さ、流暢さの評価
  - 明瞭さ: 要約が簡単に理解できるか
  - 流暢さ: 要約が言語として自然、文法誤りがないか



# 人手アノテーション

- クラウドソーシングで1つのソースに対して複数人

## ハイライトアノテーション

## ハイライトベース 内容評価

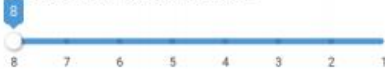
### Instructions & Controls

Your task is to **assess the quality of the summary based on the document and its highlights**.

Hover the mouse on top of ⓘ to see more information.

Words that are important in the document have been highlighted using heatmap coloring (**Darker color signifies higher importance**). You have to decide which importance level that signifies the informativeness of words.

Use the slider to remove light color (less important highlights) by sliding it to the right. The number tells you how many color you can remove until there is only one color (the most important words) left.



The 68-year-old Dutchman was appointed in March, when the Black Cats were one point above the relegation zone.

He guided them to safety and was due to leave the club in the summer, only to sign a new one-year contract.

Advocaat said: "I have made the decision to go after only eight games as I felt it was important to give everyone time to turn things around."

Sunderland chairman Ellis Short said: "I am truly saddened by Dick's decision, but I respect him for his honesty.

Media playback is not supported on this device "It is also testament to his character that he has foregone any kind of a financial settlement, something which is very unusual in football."

Assistant head coach Zeljko Petrovic has also left the club.

Saturday's draw with West Ham left Sunderland without a win in their first eight league matches and looking for a sixth manager in four years.

Since Steve Bruce was sacked in November 2011, Martin O'Neill, Paolo di Canio, Gus Poyet and Advocaat have managed the club.

Advocaat's departure also follows a trend set by Di Canio and Poyet of managers arriving to save the Black Cats from the drop, only to depart in the next season.

He initially agreed to lead Sunderland only until the end of last season, shedding tears as their survival was secured with a 0-0 draw at Arsenal.

Advocaat said he would leave the club to fulfil a promise to his wife, but changed his mind and returned in June.

The former Netherlands, Russia and South Korea boss has seen his side win just once this season - against League Two Exeter in the League Cup.

North-east rivals Newcastle are the only team below them in the Premier League table.

### Assessment

Assess the following summary.

dick advocaat has resigned as sunderland manager until the end of the season.

How strongly agree are you on the following statements?

❶ All important information is present in the summary



❷ Only important information is in the summary.



Click to submit

# Highlight-based ROUGE

- ハイライトされた重要な単語、フレーズに高い重み

$$\text{HR}_{\text{rec}}^n = \frac{\sum_{g \in n\text{-gram}(\mathcal{S})} \beta_g^n \text{count}(g, \mathcal{D} \cap \mathcal{S})}{\sum_{g \in n\text{-gram}(\mathcal{D})} \beta_g^n \text{count}(g, \mathcal{D})}$$

$$\text{HR}_{\text{pre}}^n = \frac{\sum_{g \in n\text{-gram}(\mathcal{S})} \beta_g^n \text{count}(g, \mathcal{D} \cap \mathcal{S})}{\sum_{g \in n\text{-gram}(\mathcal{S})} \text{count}(g, \mathcal{S})}$$

Precisionで分母に重みをかけると、ソースに出てこない単語が無視されてしまうため、分母の重みは1にする

$$\beta_g^n = \frac{\sum_{i=1}^{m-(n-1)} \left[ \frac{\sum_{j=i}^{i+n-1} \frac{\text{NumH}(w_j)}{\mathcal{N}}}{n} \right]_{w_{i:i+n-1}=g}}{\sum_{i=1}^{m-(n-1)} [1]_{w_{i:i+n-1}=g}}$$

単語 $w_j$ がハイライトされた回数

アノテータの数

# Highlight-based ROUGEでの評価

- ハイライトあるなしで傾向は同じ
  - Reference > TCONVS2S > PTGEN
- ハイライトありの方がTCONVS2SとPTGENで差が大
  - Prec: 6.48 v.s. 3.98, Rec: 5.54 v.s. 1.83
- 正解を使った評価ではRecallでTCONVS2SよりPTGENが上回る

Model	Highlight -based		Non High- light-based		Reference -based	
	Prec	Rec	Prec	Rec	Prec	Rec
TCONVS2S	57.42	49.95	52.55	41.04	46.75	36.45
PTGEN	50.94	44.41	48.57	39.21	44.24	38.24
Reference	67.90	56.83	66.01	52.45	—	—

# 定性的分析

## ARTICLE:

The yellow warning will remain in force until 11:00 on Sunday. Forecasters said showers accompanied by widespread sub-zero temperatures would see ice form on many untreated roads. Some snow is expected even at low levels in northern Scotland and other areas could see 2-3cm fall on higher ground. A Met office forecaster said : " Over northern Scotland showers will fall as snow to low levels. " Elsewhere within the warning area these showers will be turning increasingly wintry , with the main snow level down to between 100 and 200m by the end of the night. " Locally 2 or 3 cm of snow is possible above 200m . "

正解を使った評価だとシステムが生成した単語をあまり含んでおらず、妥当なスコアを付けられない

ハイライトベースは、人が重要とした単語を含むかどうかで評価できるため、妥当なスコアを付与

## SUMMARY:

Reference: a weather warning has been issued for most parts of scotland , with drivers urged to be aware of a risk of ice and snow .

TCONVS2S: the **met office** has issued a **yellow** `` be aware " warning for snow in parts of **northern** scotland .

PTGEN: **forecasters** have warned of severe thunderstorms across parts of scotland and scotland as snow is forecast to affect **wintry** weather .