ACL2019 参加報告

磯沼 大 / 東京大学大学院工学系研究科・博士課程

2019年9月28日

自己紹介

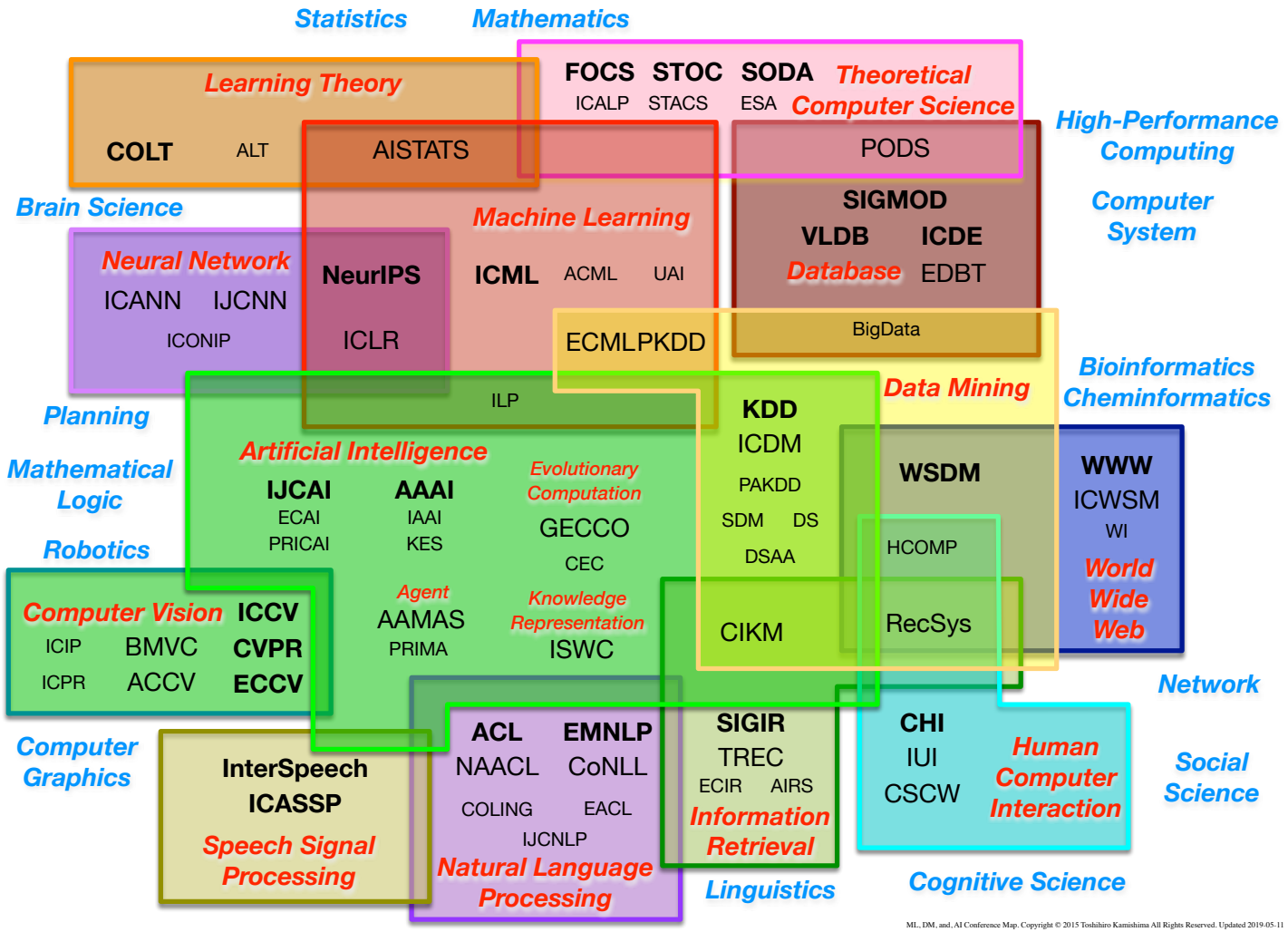
- 磯沼 大/Masaru ISONUMA
  - 東京大学大学院工学系研究科・博士課程1年
  - 坂田・森研究室
- 研究内容：半教師あり・教師なし文書要約
  - 文書分類とのマルチタスク学習を用いた重要文抽出[1)]
  - 潜在的な談話構造を捉えた商品レビューの教師なし要約生成[2)]
  - 10月からACT-X「数理・情報のフロンティア」にて
    文書の階層構造に着目した教師なし文書要約生成の研究に従事

1) M. Isonuma, T. Fujino, J. Mori, Y. Matsuo, I. Sakata. "Extractive Summarization using Multi-Task Learning with Document Classification." EMNLP (long paper), 2017.

2) M. Isonuma, J. Mori, I. Sakata. "Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking." ACL (long paper), 2019.

ACLとは

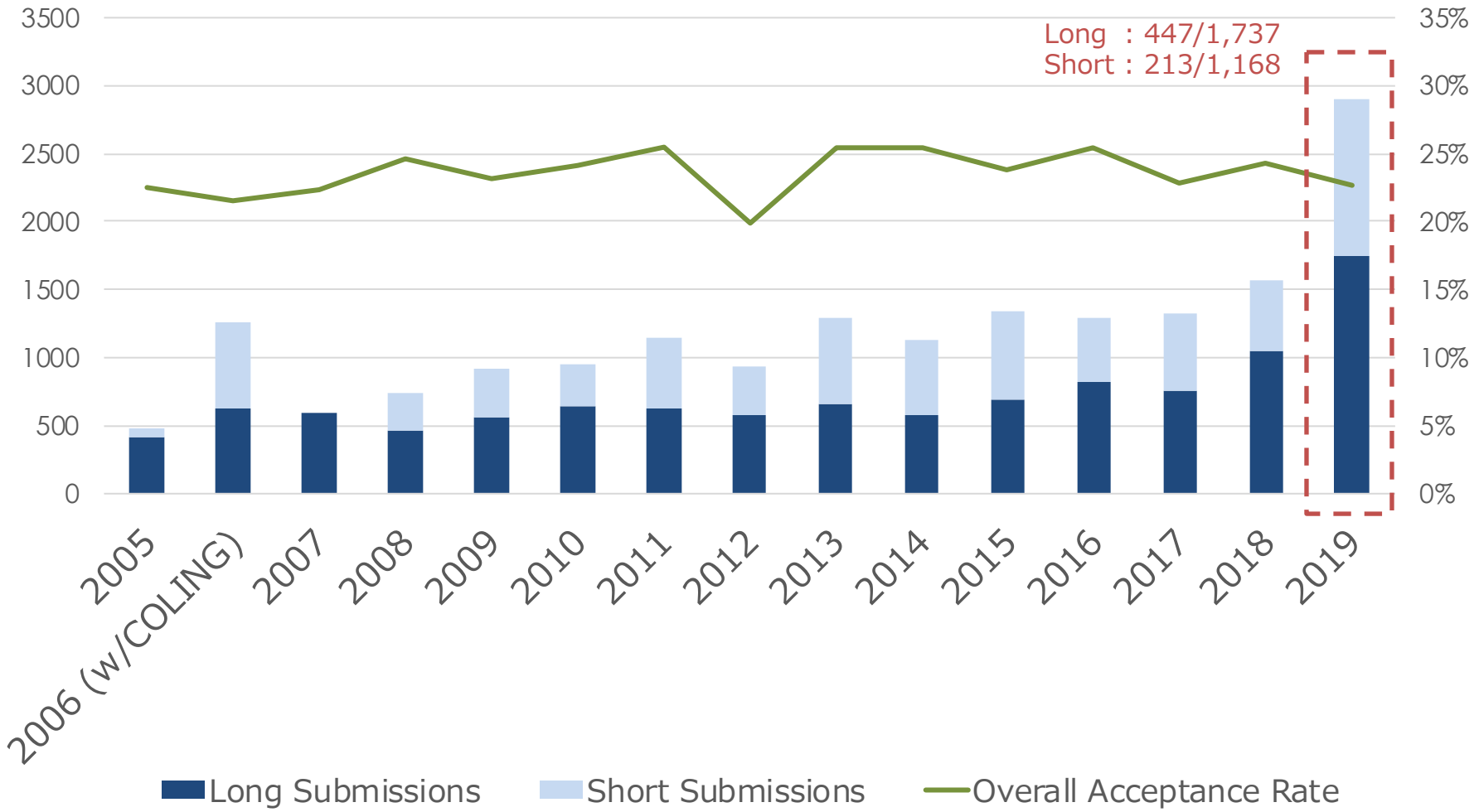
ACL: Annual Meeting of the Association for Computational Linguistics
計算言語学・自然言語処理領域のトップ会議。今年はフィレンツェで開催



出典: ML, DM, and AI Conference Map (産総研 神嶌 敏弘先生HP), http://www.kamishima.net/archive/MLDMAImap.pdf

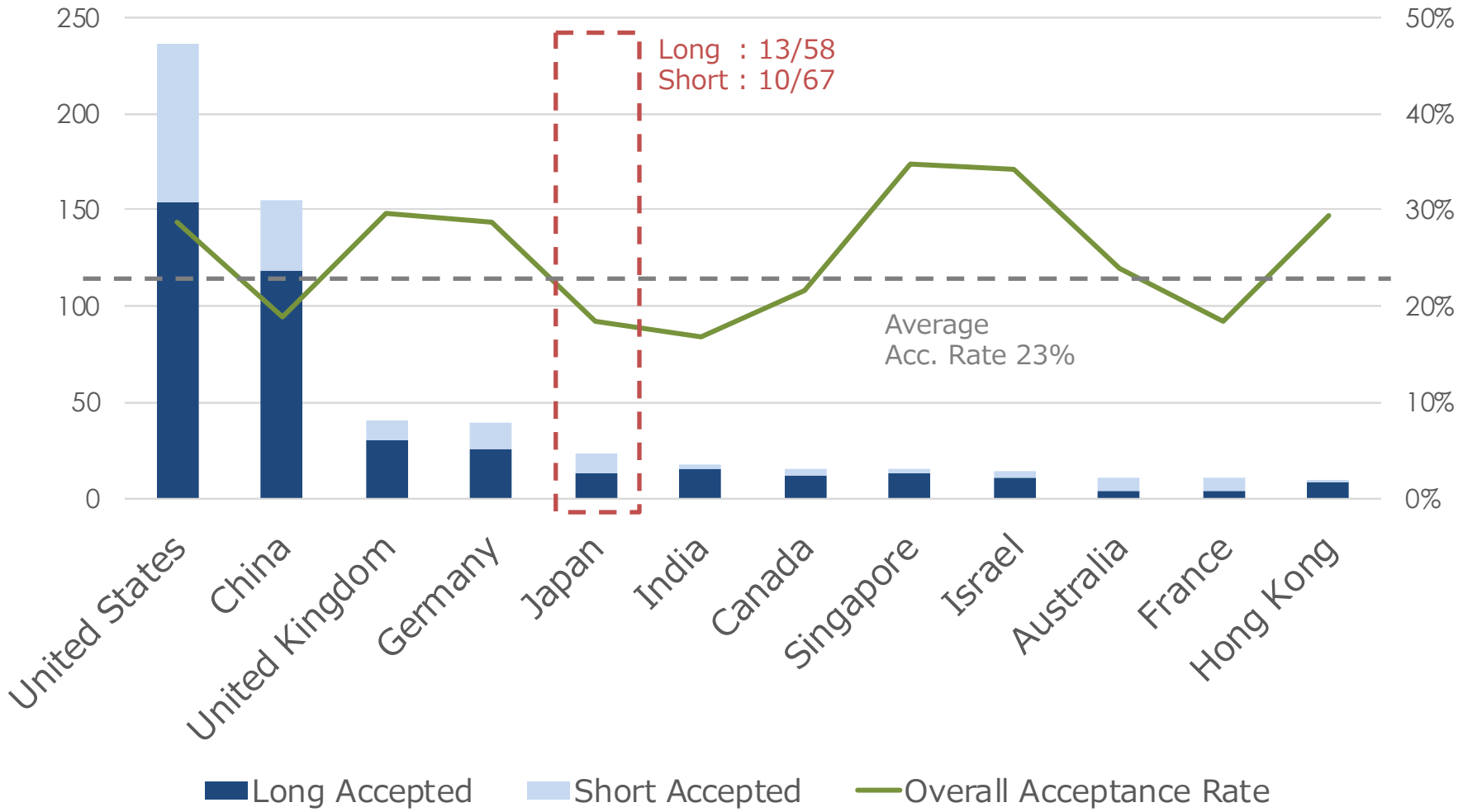
投稿数・採択率推移

今年は投稿数が昨年比75%増。採択率はほぼ変わらず



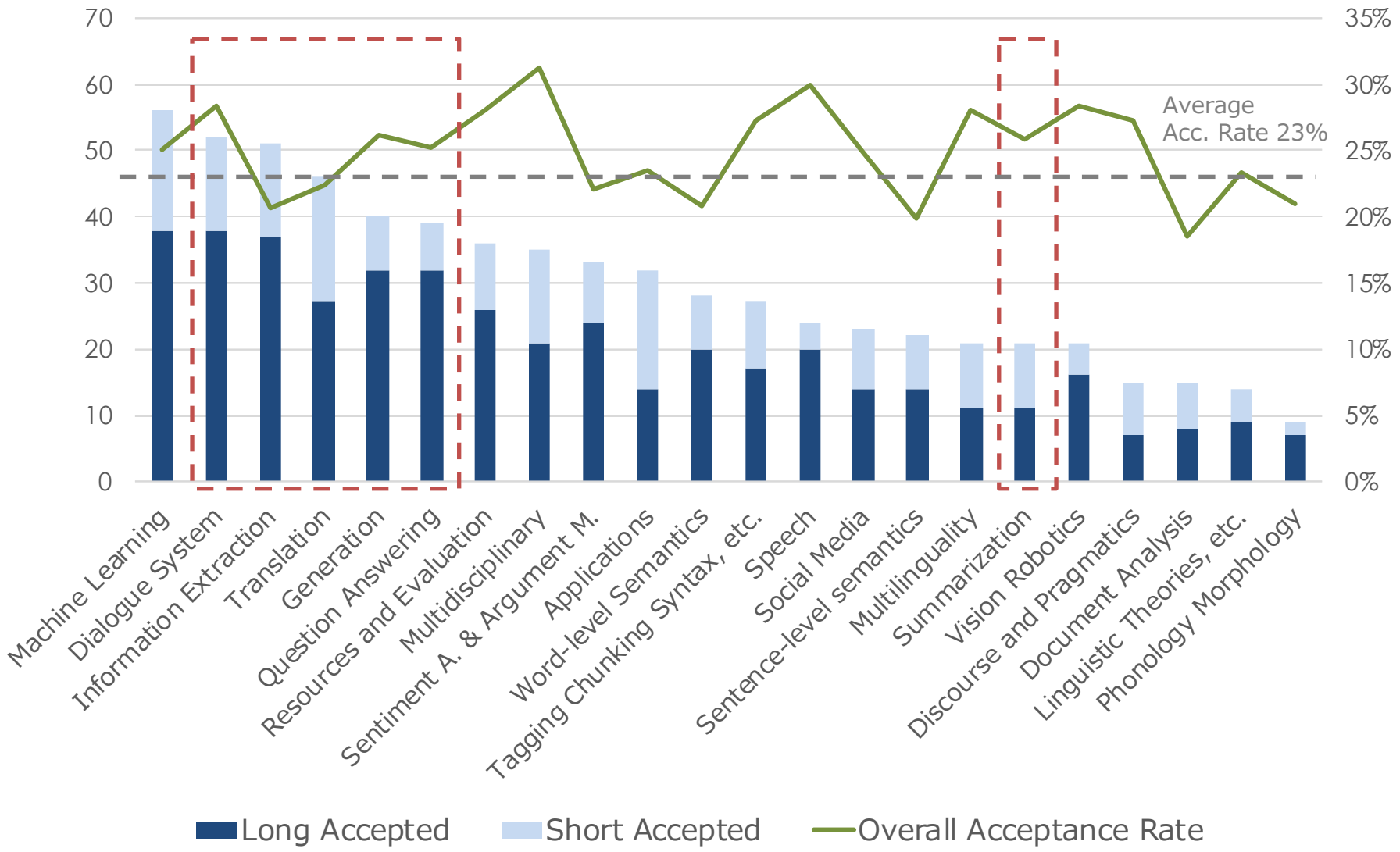
国別採択数・採択率

米中の2強。日本は5番目に採択数が多く、Shortが多い傾向



分野別採択数・採択率

対話/情報抽出/翻訳/生成/QAに対し、要約は（意外にも）少ない



スポンサーの変遷

9年前は現地の研究機関が主だったが、現在は米・中IT大手が主

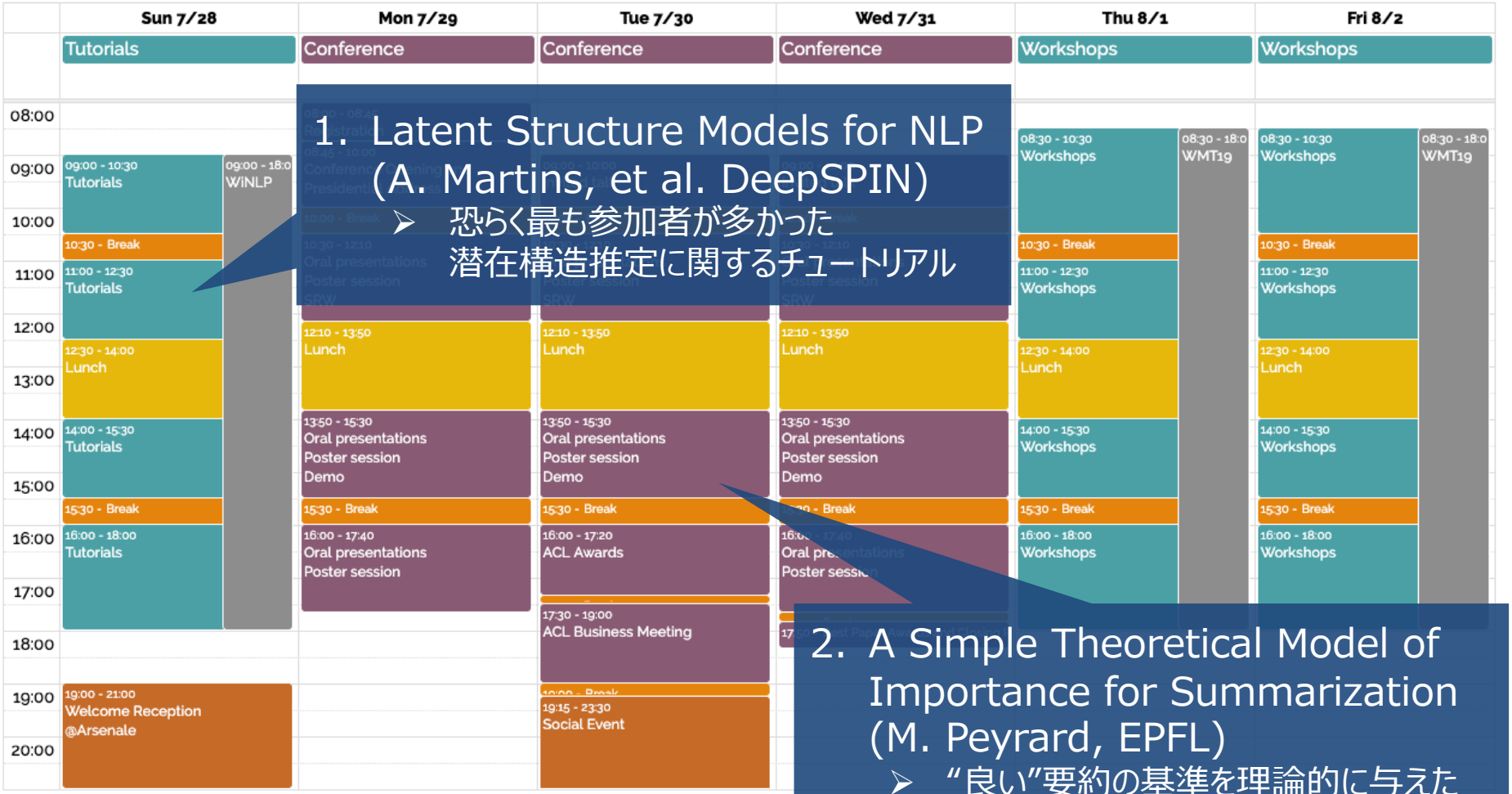
2010 (Sweden)

2014 (U.S.)

2019 (Italy)

会議スケジュールと紹介内容

個人的に興味がある・非常にレベルが高かった発表2件を紹介



1. Latent Structure Models for NLP (A. Martins, et al. DeepSPIN)
   - 恐らく最も参加者が多かった潜在構造推定に関するチュートリアル

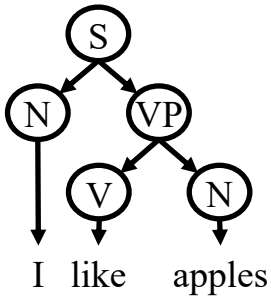
2. A Simple Theoretical Model of Importance for Summarization (M. Peyrard, EPFL)
   - “良い”要約の基準を理論的に与えたOutstanding Paper

出典: ACL2019HP, http://www.acl2019.org/EN/program.xhtml

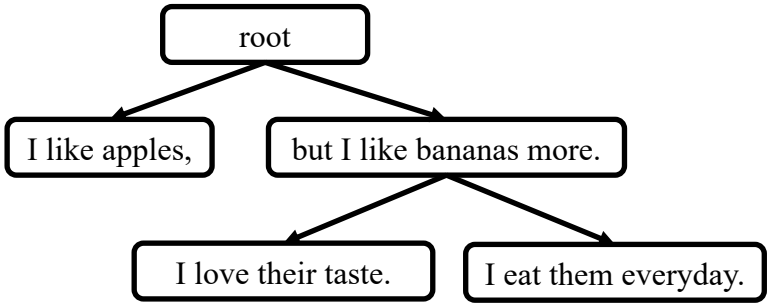
Latent Structure Models for NLP

潜在構造推定のチュートリアル (A. Martins, et al. DeepSPIN)

- 構造推定：文・文書などに内在する構造を推定するタスク
  - 構造推定は、下流タスク（分類、情報抽出、翻訳など）を解く上で有用



句構造木



談話依存構造木

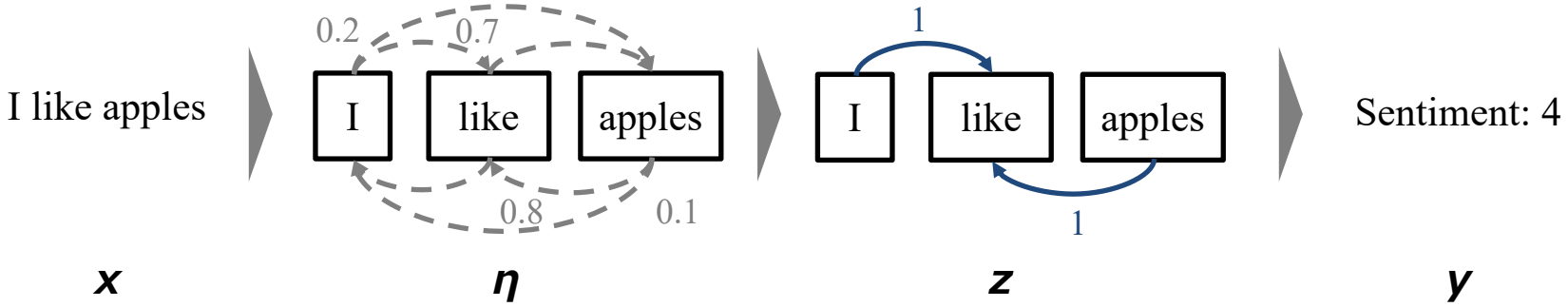
などなど…

- これまでは構造推定タスクを個別に解いてから下流タスクを解くのが主流
  - 利点：高速・シンプル、各タスクのモジュールを入替可能
  - 欠点：構造のアノテーションが必要、構造推定タスクのエラーが下流タスクに伝播

潜在構造推定のモチベーション

構造を潜在変数とみなし下流タスクを解く研究が盛んに

- モデルの解釈性が向上、構造に関する事前知識を導入可能、
  （願わくば）少量のパラメータで高い予測精度が得られる

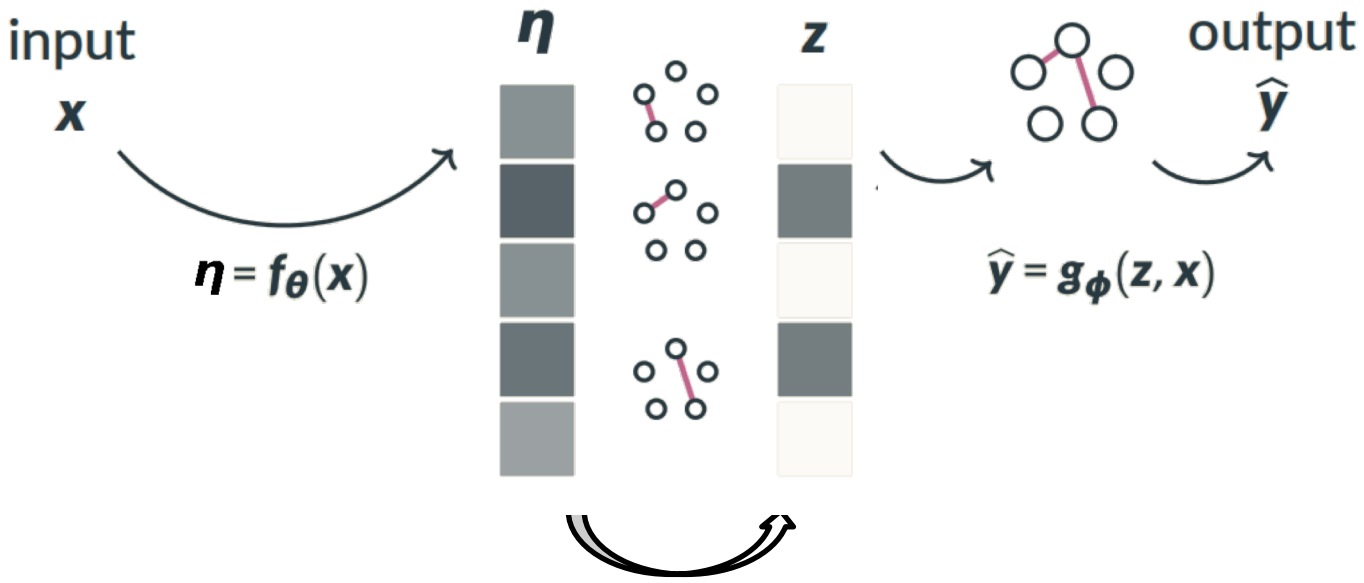


- 色々な研究あるけど正直よく分からん！ → アプローチを系統立てて解説
  1. Straight Through Estimator
     - Y. Bengio, et al. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. CoRR, 2013.
     - C. Corro & I. Titov. Learning Latent Trees with Stochastic Perturbations and Differentiable Dynamic Programming. ACL, 2019.
  2. REINFORCE
     - D. Yogatama, et al. Learning to compose words into sentences with reinforcement learning. ICLR, 2017.
  3. Structured Attention
     - Y. Kim, et al. Structured Attention Networks, ICLR, 2017.
     - Y. Liu and M. Lapata. Learning Structured Text Representations. TACL, 2018.

※ 深層生成モデルによるアプローチはスコープ外。EMNLP18のチュートリアル
“Deep Latent-Variable Models for Natural Language” (Kim, Wiseman, Rush) がオススメされています

潜在構造推定の枠組みの背景

構造を示す潜在変数$\boldsymbol{z}$は離散のため、学習には工夫が必要



種々のアルゴリズムで
事後確率が最大となる
構造$\boldsymbol{z}$を推定

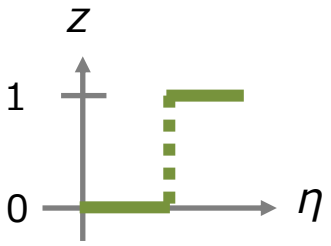
$$\boldsymbol{z} = f_{MAP}(\boldsymbol{\eta})$$

$$f_{MAP} \in \{Viterbi, CKY, Max.SpanningTree, etc.\}$$

$\boldsymbol{z}$を用いて下流タスクの
損失を最小化したいが…

$$L(\mathbf{z}) = L(\hat{y}(\mathbf{z}; x), y))$$

But… $\frac{\delta \boldsymbol{z}}{\delta \boldsymbol{\eta}} = \mathbf{0}$ → $\boldsymbol{\theta}$を更新不可！



1. Straight-Through Estimator

Backward時に$\boldsymbol{z}$を$\boldsymbol{\eta}$で偽装することで、勾配が伝播するように

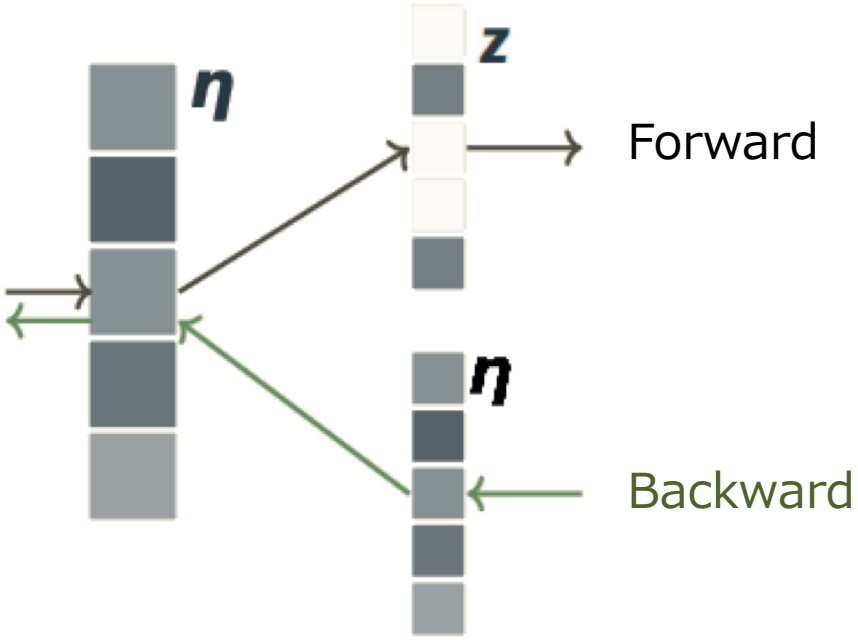
損失関数

$L(\mathbf{z})$

- Forward:
  - $\boldsymbol{z} = f_{MAP}(\boldsymbol{\eta})$
- Backward:
  - $\boldsymbol{z}$を$\boldsymbol{\eta}$で偽装 → $\frac{\delta \boldsymbol{z}}{\delta \boldsymbol{\eta}} = \mathbf{1}$ → 勾配を伝播可能

$\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}|x)}[L(\mathbf{z})]$

$L\left(\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}|x)}[\mathbf{z}]\right)$



2. REINFORCE

$\boldsymbol{z}$を確率変数とみなし、$\pi_{\boldsymbol{\theta}}(\boldsymbol{z}|x)$からサンプリング、 $\boldsymbol{\theta}$を学習

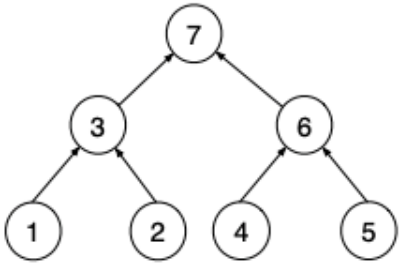
損失関数

$L(\boldsymbol{z})$

$\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\boldsymbol{z}|x)}[L(\boldsymbol{z})]$

$L(\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\boldsymbol{z}|x)}[\boldsymbol{z}])$

- $\boldsymbol{z}$を$z_j \in \{Shift, Reduce\}$のアクションの系列とみなし、$\boldsymbol{z} \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{z}|x)$によりサンプリング



$\boldsymbol{z}$ = [*Shift, Shift, Reduce, Shift, Shift, Reduce, Reduce*]

- サンプリングした$\boldsymbol{z}$に基づき、報酬(下流タスクの性能)$L(\boldsymbol{z})$を最大化

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{z} \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{z}|x)}[L(\boldsymbol{z})] = \nabla_{\boldsymbol{\theta}} \left[\sum_{\boldsymbol{z}} L(\boldsymbol{z}) \pi_{\boldsymbol{\theta}}(\boldsymbol{z} \mid x) \right]$$

$$= \mathbb{E}_{\boldsymbol{z} \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{z}|x)}[L(\boldsymbol{z}) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{z} \mid x)]$$

→ 勾配が伝播し、$\boldsymbol{\theta}$を更新可能

3. Structured Attention

zそのものではなく、$\pi(z)$ を推定し、そのまま下流タスクで利用

損失関数

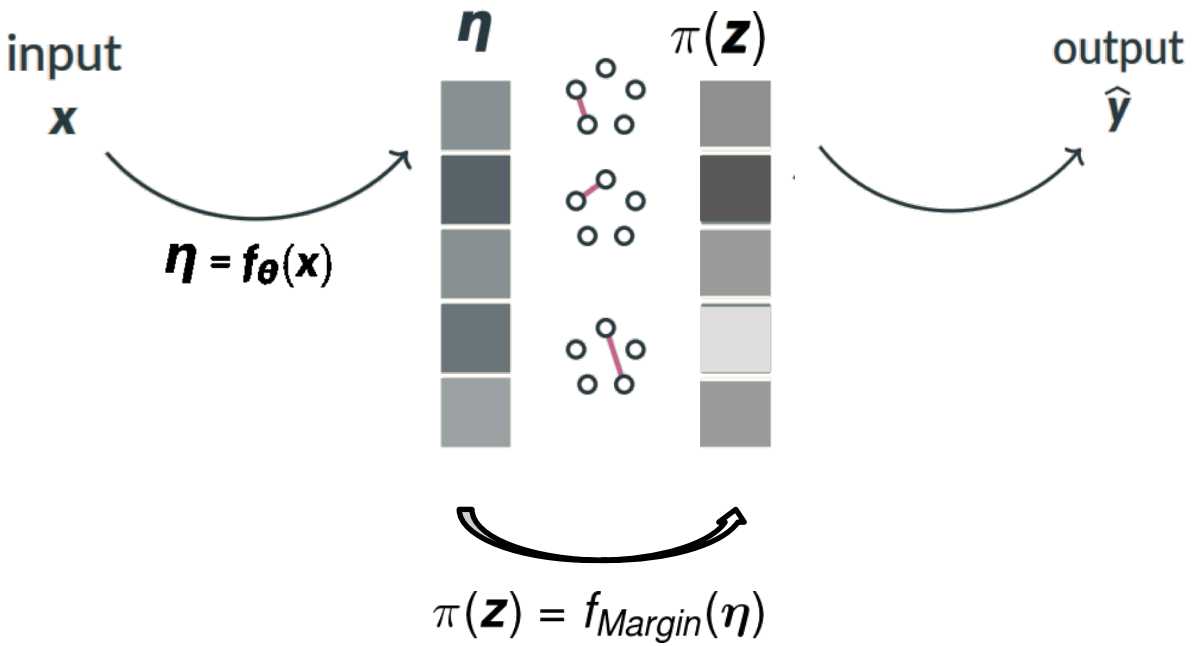
$L(z)$

$\mathbb{E}_{\pi_\theta(z|x)}[L(z)]$

$L(\mathbb{E}_{\pi_\theta(z|x)}[z])$

- pair-wise attentionで得たηを、前向き後向きアルゴリズムや行列木定理により、zの確率 $\pi(z)$ に変換

→ $\pi(z)$ は連続なため、勾配を伝播可能



$f_{Margin} \in \{Forward-Backward, Inside-Outside, Maxtrix-Tree, etc.\}$

まとめ

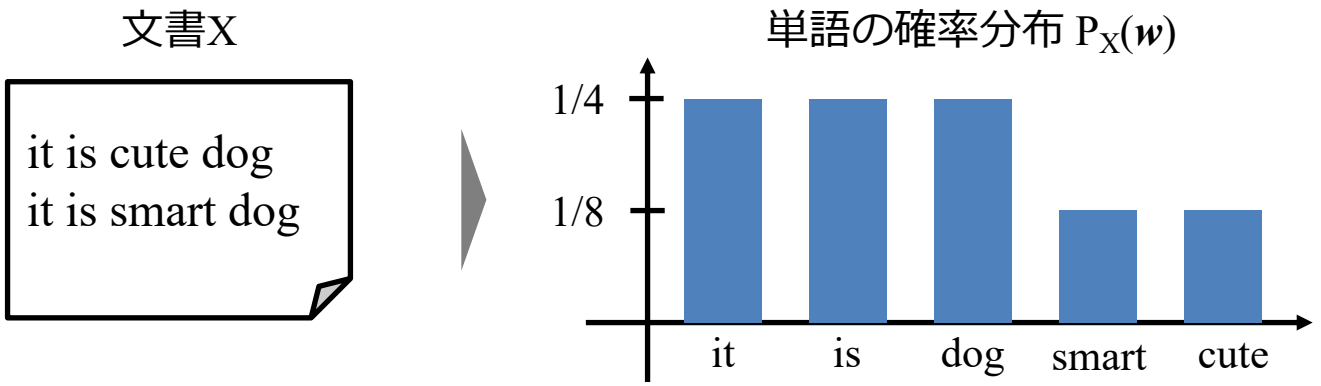
損失関数	
$L(\mathbf{z})$	1. Straight Through Estimator ・ 利点：シンプル、計算量が効率的 ・ 欠点：backwardとforwardの不適合により、勾配にバイアスが生じる
$\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}\|x)}[L(\mathbf{z})]$	2. REINFORCE ・ 利点：勾配にバイアスが生じない ・ 欠点：アクション（構造）の数が膨大で、構造毎に勾配が得られるため、勾配の分散が大きい
$L(\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}\|x)}[\mathbf{z}])$	3. Structured Attention ・ 利点：勾配にバイアスが生じない、勾配の分散が小さい ・ 欠点：各構造の確率が計算できない場合は適用不可、ケース毎に効率的な計算アルゴリズムを組む必要

A Simple Theoretical Model of Importance for Summarization

“良い”要約のスコア関数を提案 (M. Peyrard, EPFL)

- 多くの要約モデルは見本をもとに”良い”要約を学習するアプローチが多く、その基準は明示的に考慮されていない
  - 一部の基準はモデルに組み込まれているものの、統合的にそれらを扱えていない

- ”良い”要約の基準と、それを統合したスコア関数を提案
  - 文書を単語の確率分布と見立てる



  - 要約の単語分布の”良さ”を、3つの基準を統合してスコアリング
    - 要約の冗長性
    - 元の文書に対する関連性
    - 背景知識に対する新規性

基準1: 冗長性

基準

Redundancy
（冗長性）

Relevance
（関連性）

Informativeness
（新規性）

- ”良い”要約は冗長でない（多くの情報を持つ）
  - 要約Sの冗長性は平均情報量$H(S)$で表せる

$$Red(S) = -H(S)$$
$$= \sum_{\omega_i} \mathbb{P}_S(\omega_i) \cdot \log(\mathbb{P}_S(\omega_i))$$

  - 全ての単語が一度のみ出現する場合に$Red(S)$は最小に
    - S_1) dog dog : 0
    - S_2) cute dog : -log(2)
    - S_3) cute cute : 0

- 劣モジュラ性を用いた最大被覆問題の背景となる考え

基準2: 関連性

基準

Redundancy
（冗長性）

Relevance
（関連性）

Informativeness
（新規性）

- ”良い”要約は元の文書とよく関連する
  - 元の文書Dに対する要約Sの関連度合いは
    クロスエントロピー$CE(S, D)$を用いて表せる

$$Rel(S, D) = -CE(S, D)$$

$$= \sum_{\omega_i} \mathbb{P}_S(\omega_i) \cdot \log(\mathbb{P}_D(\omega_i))$$

  - 元の文書に頻出する単語を含むほど$Rel(S, D)$は増大

D) it is cute dog
it is smart dog

- S_1) it is cute : -7/3 log(2)
- S_2) it is dog : -2 log(2)

∧

- 様々な要約モデルがこの仮定を背景に
  - 頻出単語/n-gramを含む文を抽出 (SumBasic, ICSI, etc.)
  - トピックモデルベース
  - グラフベース（LexRank, etc.)

基準3: 新規性

基準

Redundancy
（冗長性）

Relevance
（関連性）

Informativeness
（新規性）

- ”良い”要約は背景知識にない情報を持つ
  - 背景知識Kに対する要約Sの新規性は
    クロスエントロピー$CE(S, K)$を用いて表せる

$$Inf(S, K) = CE(S, K)$$
$$= -\sum_{\omega_i} \mathbb{P}_S(\omega_i) \cdot \log(\mathbb{P}_K(\omega_i))$$

Rel(*S*, *D*)と
符号が逆

  - 背景知識に頻出する単語を含むほど*Inf*(*S*, *K*)は減少

K) it is cute dog
   it is smart dog

- S_1) it is cute : 7/3 log(2)
- S_2) it is dog : 2 log(2)

- 既存研究では比較的考慮されてこなかった
  - idf ベース
  - update summarization (TAC-2008)

3基準を統合した要約のスコア

- 元の文書D, 背景知識Kが与えられた時の要約Sのスコア$\theta_I(S, D, K)$を、3つの基準を統合して定義

$$\theta_I(S, D, K) \equiv -Red(S) + \alpha Rel(S, D) + \beta Inf(S, K)$$

- θ_Iは以下のように解釈できる

$$\mathbb{P}_{\frac{D}{K}}(\omega_i) = \frac{1}{C} \cdot \frac{d_i^\alpha}{k_i^\beta}$$

$$C = \sum_i \frac{d_i^\alpha}{k_i^\beta}, \quad \alpha, \beta \in \mathbb{R}^+$$

$$d_i = \mathbb{P}_D(\omega_i) \quad k_i = \mathbb{P}_K(\omega_i)$$

単語w_iの重要性

- 元の文書Dに頻出するほど重要性は高い
- 背景知識Kに頻出するほど重要性は低い

$$\theta_I(S, D, K) = -KL(\mathbb{P}_S, ||\mathbb{P}_{\frac{D}{K}})$$

重要性分布に近い単語の分布を持つ要約ほど高スコア

実験設定・結果

提案スコアは人手評価と高い相関を示す

実験設定

- 評価方法
  - TAC shared Task (複数文書要約)で既に人手評価された要約と、提案法による評価との相関を計算
  - Generic: 10文書を要約（背景知識$P_K(w)$は一様分布）
  - Update: 最初の10文書を背景知識として扱い、別の10文書を要約
- データセット
  - TAC-2008, 2009
- ベースライン
  - ICSI, LexRankなどによる文のスコアリング

人手評価との相関

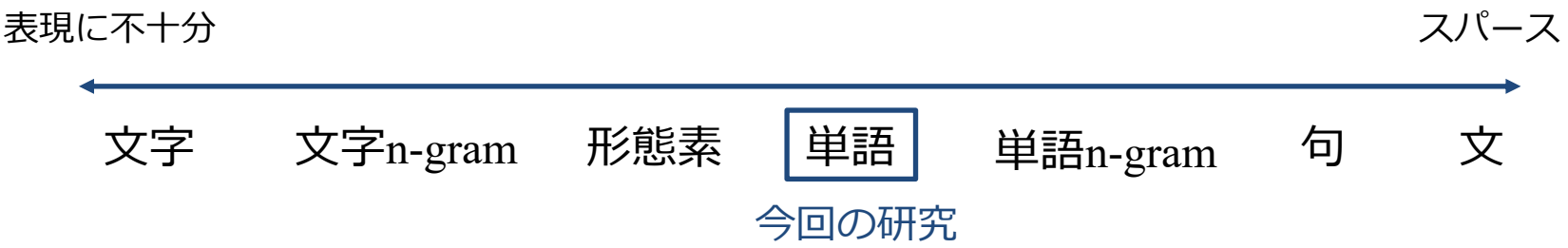
	Generic	Update
ICSI	.178	.139
Edm.	.215	.205
LexRank	.201	.164
KL	.204	.176
JS	.225	.189
KL_{back}	.110	.167
JS_{back}	.066	.187
Red	.098	.096
Rel	.212	.192
Inf	.091	.086
θ_I	**.294**	**.211**

- 特にGenericにおいて、提案スコアが人手評価と高い相関を示す

将来展望

- 提案スコア関数を種々の最適化手法と組み合わせることで、
  より人手に近い要約を生成できることを期待

- 意味単位をどう設定するか？



- 背景知識をどう設定するか？
  - 参照要約を用いることで、背景知識Kやα, βなどのハイパーパラメータを探索
  - ドメイン単位、ユーザ単位での背景知識設定も考えられる