COLING2016 参加報告

2016/12/21

Shoko Suzuki

IBM Research - Tokyo

Overall review of COLING2016@OSAKA

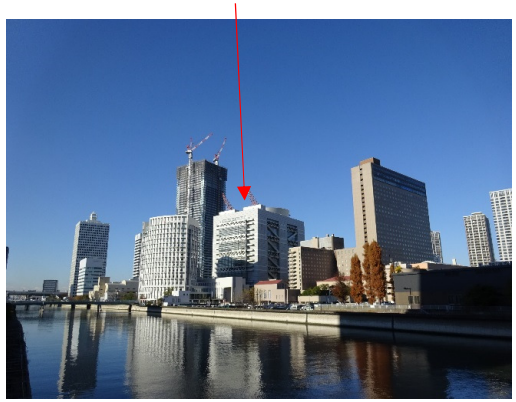
- 2年に一度開催
- 3.5day main conference + 2day workshop + tutorials + half-day excursion
  - Main conferenceでは、4 oral sessions + 1 poster session + demo session が並行
- 1039 submissions (reviewed)
- 337 Accepted papers
  - Oral :135
  - Poster:202
- Acceptance rate:
  - Oral + poster :32%

Overall review of COLING2016@OSAKA

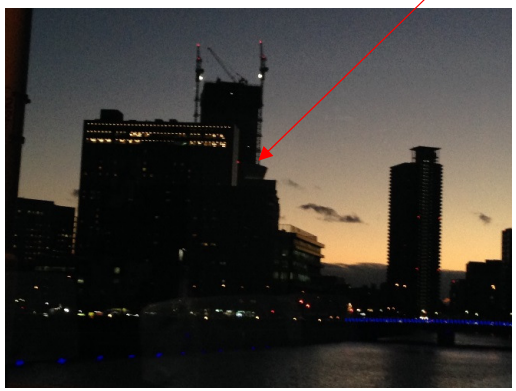
- 国際色豊か
- 特に中国からの投稿が多い
  - Accepted papers 337件のうち76件が中国。
  - US 56件、ドイツ 39件、UK 26件、日本 19件、インド 17件、、、
- 参加者1200人超は過去最高

- NNをなんらかの形で利用したものが多い
- 一方で、他分野との境界領域のような話も多い

会場



このあたりが会場



Invited talks

- “Universal Dependencies - Dubious Linguistics and Crappy Parsing?”
  - Joakim Nivre
- “Getting the Input Right: Refining our Understanding of What Children Hear”
  - Reiko Mazuka
- “NLP to support clinical tasks and decisions”
  - Dina Demner-Fushman
- “A Look at Computational Argumentation and Summarisation from a Text-Understanding Perspective”
  - Simone Teufel

“D-GloVe: A Feasible Least Squares Model for Estimating Word Embedding Densities”(1/3)

- 目的：Word embeddingにおいて、densityを導入することで、co-occurrenceの信頼度が低い場合及びuninformativeな語に対処する。

- P(j|i)を $\frac{x_{ij}}{\sum_l x_{il}}$ で推定してよいのは、iの頻度が高い場合のみ。
  - x_{ij} ：jがiのcontextの元で出現する頻度
- Dirichlet-Multinomial modelを仮定し、よりロバストにP(j|i)の推定を行った。

“D-GloVe: A Feasible Least Squares Model for Estimating Word Embedding Densities”(2/3)

- uninformativeな場合の影響を小さくしたい。
  - $w_i \cdot \tilde{w}_j$ と、前頁で推定したlogP(j|i)とがどれくらい異なるかをvarianceで評価

Frequent and informative

	Term Frequency	$Var[e_j^{info}]$
one	411764	1.39
time	21412	0.720
states	14916	0.259
united	14494	0.282
city	12275	0.221
university	10195	0.632
french	8736	0.270
two	192644	0.815
american	20477	1.26
government	11323	1.54

Frequent and uninformative

	Term Frequency	$Var[e_j^{info}]$
in	372201	58.08
was	112807	43.16
or	68945	57.10
his	62603	47.05
also	44358	44.87
their	31523	84.35
used	22737	31.80
these	19864	25.96
e	11426	45.75
without	5661	30.38

Infrequent and uninformative

	Term Frequency	$Var[e_j^{info}]$
wendell	40	29.38
actuality	42	29.75
ebne	54	30.17
christology	45	31.04
mico	45	33.38
utilised	30	21.52
reopened	54	21.32
generalizes	24	19.07
flashing	49	19.83
eitc	27	20.77

Infrequent and informative

	Term Frequency	$Var[e_j^{info}]$
psycho	56	0.05
quantization	56	0.25
residue	56	0.02
inert	54	0.98
imap	54	0.19
batsman	52	0.68
bilinear	52	0.18
crucified	50	0.08
germanium	50	0.11
lactose	50	0.45

“D-GloVe: A Feasible Least Squares Model for Estimating Word Embedding Densities”(3/3)

- Varianceの逆数のようなものを重みとしてword embeddingを導出

$$\sum_{i=1}^{n}\sum_{j\in J_i}\frac{1}{\sigma_{ij}^2}(w_i\cdot\tilde{w}_j+\tilde{b}_j-s_{ij})^2$$

理解が怪しいので、是非元論文を読んでください

- Word analogy のtask で評価
- 結果として、高頻度語ではうまくいくが、低頻度語ではSGのほうが良い状態

様々な分野の研究発表があり、裾野の広さを感じました。例えば

- エッセイの採点
  - "Evaluating Argumentative and Narrative Essays using Graphs"
  - "Using Argument Mining to Assess the Argumentation Quality of Essays“
- 科学論文の長期スパンでの変遷
  - “Modeling Diachronic Change in Scientific Writing with Information Density”
- 時間関係の分類
  - "On the contribution of word embeddings to temporal relation classification“
- 特許請求項の解析 <-自分の話
  - “Extraction of Keywords of Novelties from Patent Claims”
- Computational Psycholinguistics が一つのセッション

“Using Argument Mining to Assess the Argumentation Quality of Essays“

- エッセイの各センテンスをタイプに分類

Paragraph	Premise: *Secondly, most violent crimes are related to the abuse of guns, especially in some countries where guns are available for people.* Conclusion: *Eventually, guns will create a violent society if the trend continues.* Premise: *Take an example, in American, young adults and even juveniles can get access to guns, which leads to the tragedies of school gun shooting.* Premise: *What is worse, some terrorists are able to possess more advanced weapons than the police, which makes citizens always live in danger.*
ADU flow	(1x Premise, 1x Conclusion, 2x Premise)
ADU change flow	(Premise, Conclusion, Premise)

- タイプの並びがスコア予測に役立つことを示した

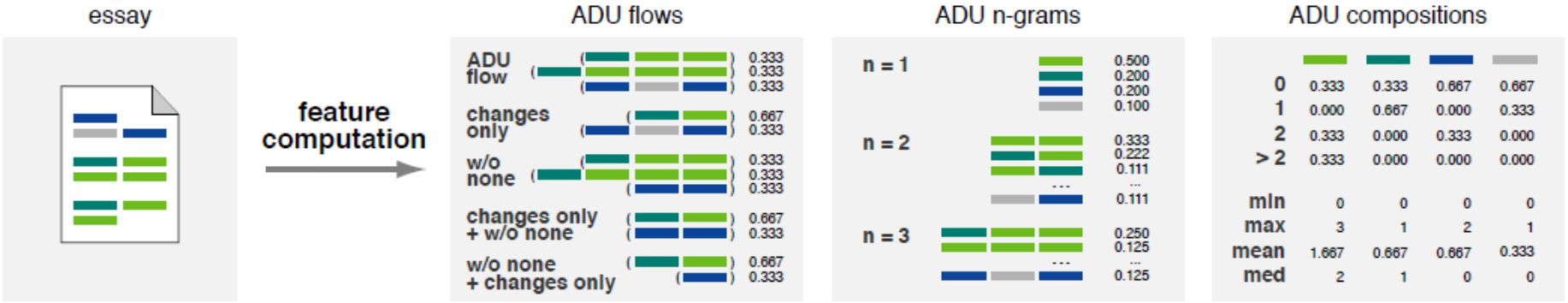


Figure 4: Sketch of the three feature types that we propose based on the output of argument mining.

“Modeling Diachronic Change in Scientific Writing with Information Density”

- １９世紀と20世紀の論文が、どのように異なっているかを分類タスクで評価
- 単語の表層レベルで異なるのは明らかだが、POS n-gramでも大きな違いがある
- 20世紀のPOS n-gramは、長いがある程度パターンは決まっている

	POS 3-gram	examples
19th century	, IN DT	*, that the, , in the, , on the*
	IN DT NN	*by the author, in this paper, of the earth*
	NN , CC	*water , and, acid , and, light , and*
	DT NN ,	*the author, , this paper, , the earth ,*
	DT NN IN	*the action of, the quantity of, the surface of*
20th century	JJ NN NN	*superficial gas velocity, natural language processing, optimal control problem*
	DT NN NN	*a computer program, the heat transfer, the nucleotide sequence*
	NN NN IN	*nucleotide sequence of, sequence analysis of, gene expression in*
	JJ NN NNS	*partial differential equations, open reading frames, linear matrix inequalities*
	NN NN .	*gene expression ., power consumption ., control system .*

Table 3: Most frequent lexical realizations of top 5 POS 3-grams for 20th c. and 19th c. abstracts

“On the contribution of word embeddings to temporal relation classification"

- 目的: 文書から、事象の前後関係を抽出することを試みる

> *When Wong Kwan* ***spent*** $_{e_1}$ *seventy million dollars for this house, he thought it was a great deal. He* ***sold*** $_{e_2}$ *the property to five buyers and said he'd double his money.*

- 文書から事象関係の特徴量を抽出し、分類
- Global な featureとしてword embeddingを利用
- Word embeddingと分類された特徴量の両方を使うことで、前後関係の抽出精度が向上


ありがとうございました