ACL2017参加報告(2)
～マルチモーダルな自然言語処理～

村岡雅康
IBM Research – Tokyo
日本IBM
2017/09/08

Vancouver aquarium

マルチモーダル研究の高まり

- マルチモーダル研究とは複数のモダリティを扱う研究
  - 画像のキャプション生成，音声のspeech to text等
- マルチモーダル関連の論文数
  - ACL2016：
    - 1 tutorial (“Multimodal Learning and Reasoning”)
    - 1 session (“Language and Vision”)
    - 10 papers
  - ACL2017：
    - 1 tutorial (“Multimodal Machine Learning”)
    - 1 Invited Talk (by Mirella Lapata)
    - 2 session (“Vision I&II”)
    - 16 papers

紹介する論文

- “Learning Character-level Compositionality with Visual Features”
  - Frederick Liu, Han Lu, Chien Lo, and Graham Neubig
- “Learning word-like units from joint audio-visual analysis”
  - David Harwath and James Glass
- “FOIL it! Find One mismatch between Image and language caption”
  - R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, and R. Bernardi
- “On the Challenges of Translating NLP Research into Commercial Products”
  - Daniel Dahlmeier

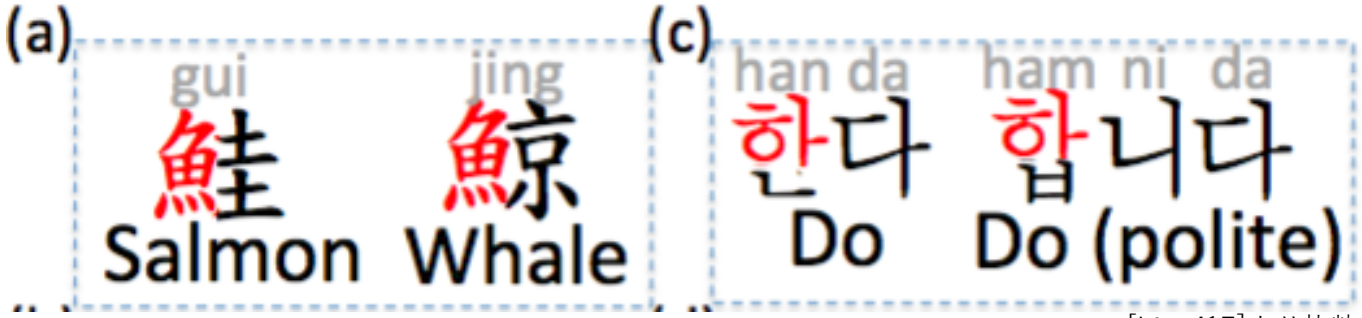

Vancouver Harbour

Learning Character-level Compositionality with Visual Features

Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig

文字embeddingを文字の画像から学習する話

- 単語のembeddingは共起文脈をつかって学習
- レアな単語のembeddingは学習が難しい
  →文字embeddingを学習
- 文字の種類が多い言語(中国語，日本語，韓国語)はレアな文字(例：漢字)の学習が難しい
  →**文字の画像から**embeddingを学習
  - 部首など文字の一部は文字間で共通している
  - 同じ部分を持つ文字同士の意味は似ている



[Liu+'17]より抜粋

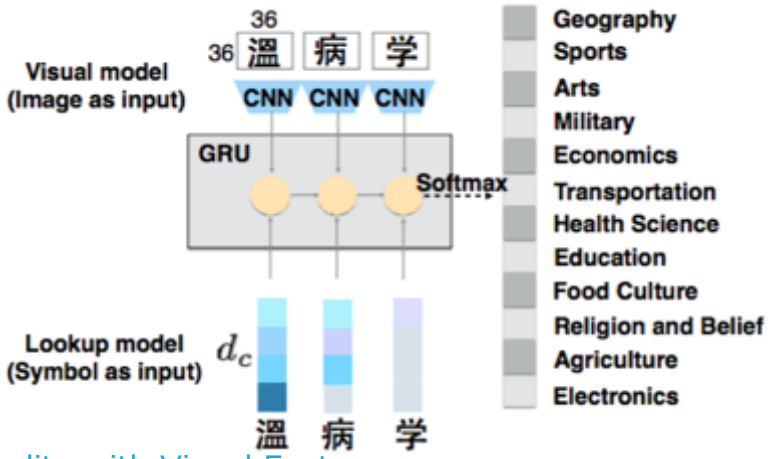
[P17-1188] Learning Character-level Compositionality with Visual Features

実験

- 問題設定：Wikipediaのタイトルからその記事のカテゴリ(全12種類)を予測する分類問題
  - 対象言語：中国語(traditional/simplified)，日本語，韓国語
- ベースライン：一般的な文字embedding
- 結果：低頻度文字では提案手法が良い
  - **文字間で共通部分を共有している**ことを示唆

	Rank		
Lookup/Visual	100	1000	10000
zh_trad	0.22/**0.49**	0.35/0.35	**0.40**/0.39
zh_simp	0.25/**0.53**	**0.39**/0.37	**0.41**/0.40
ja	0.30/**0.35**	**0.45**/0.41	**0.44**/0.41
ko	**0.44**/0.33	**0.44**/0.33	**0.48**/0.42

[Liu+'17]より抜粋

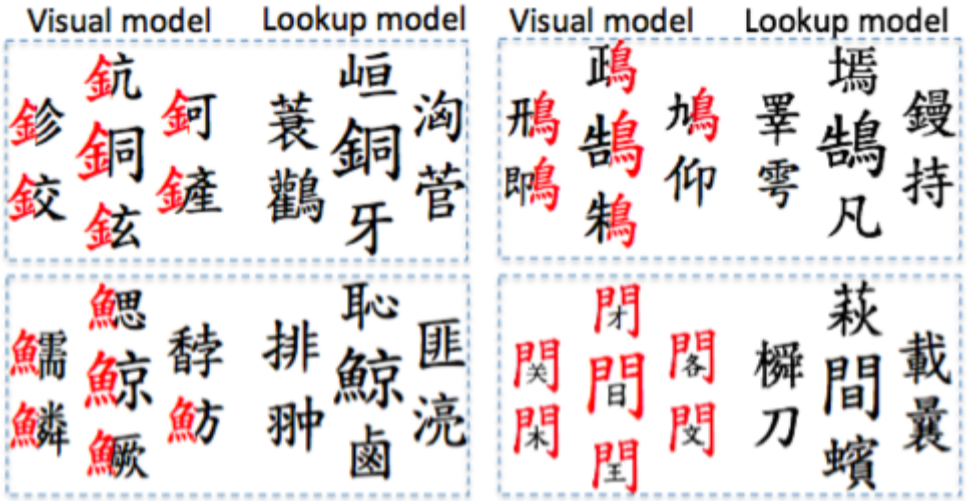


[P17-1188] Learning Character-level Compositionality with Visual Features

可視化

- 文字のどの部分が分類に寄与しているかを可視化
  - 色が濃いほど寄与度大
  - モデルは**部首に注目するように学習**

- Nearest neighbors
  - **共通した部分を持つ文字同士近くに分布**



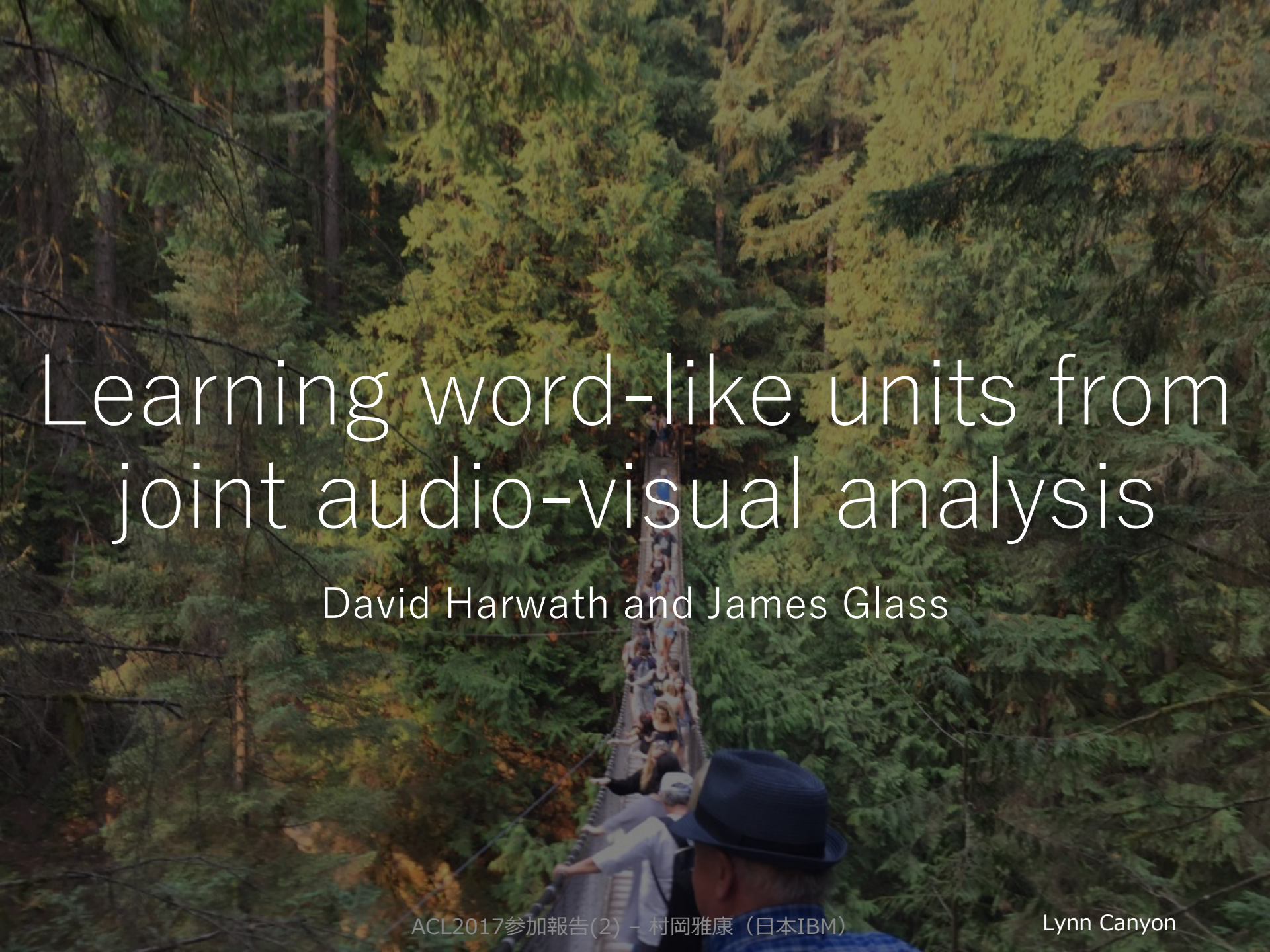
[Liu+'17]より抜粋

[P17-1188] Learning Character-level Compositionality with Visual Features

所感

- 画像を使って文書(タイトル)分類を行う面白い試み
  - あるモダリティ(画像)が他のモダリティ(言語)を補完する良い例
- まだ文字の大きな属性しか捉えられてなさそう
  - semanticな情報までembeddingに含まれてない？
  - 部首レベルでking – man + woman ~ queenみたいなことはまだできなさそう
- これをSkip-gram[Mikolov+'13]で学習したら面白そう
  - なぜベースラインとしてpre-trainedのSkip-gramを使っていないのか疑問

[P17-1188] Learning Character-level Compositionality with Visual Features



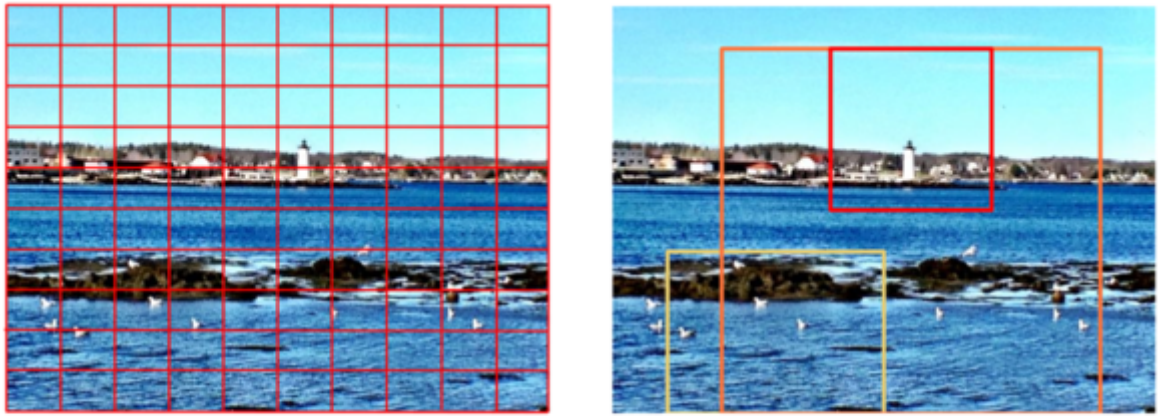
Learning word-like units from joint audio-visual analysis

David Harwath and James Glass

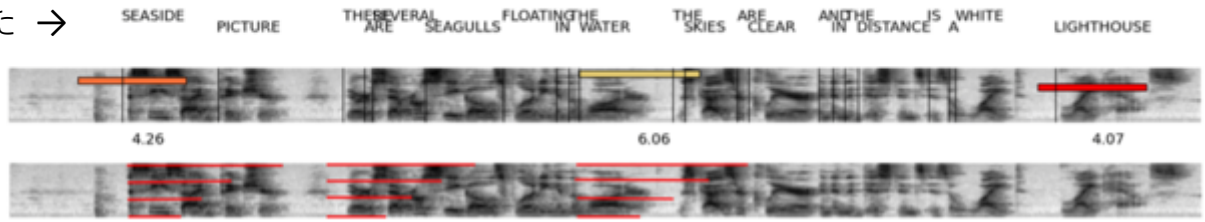
Lynn Canyon

画像と音声のパラレルコーパスから テキストデータを使わずに同じ概念を アラインする話

- 画像のregionと音声のintervalのアライン
  - モデルの学習は1画像全体と対応するキャプション(音声)全体のペアのみで行っているだけ！
  - アライン時には物体認識も音声認識も行っていない



音声を
自動認識させた →
テキスト
(アライン時は
使っていない)



[Harwath+'17]より抜粋

[P17-1047] Learning word-like units from joint audio-visual analysis

実験

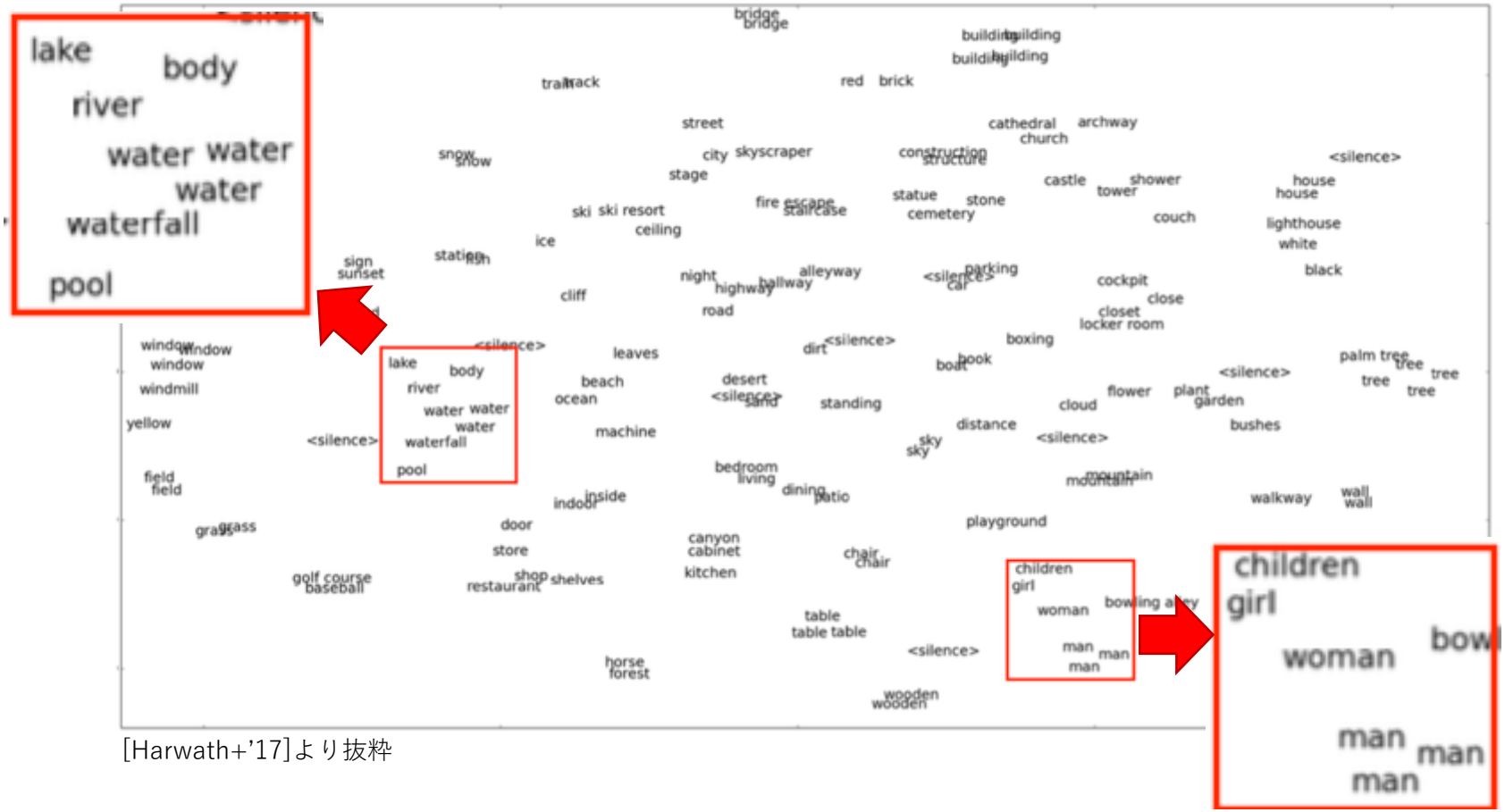
- 定量評価もしているがあまり面白くないので割愛
- 定性分析：アラインされた画像と音声のペア

[Harwath+'17]より抜粋

[P17-1047] Learning word-like units from joint audio-visual analysis

音声の“word-like units”の可視化

- 似た意味のembeddingが近くに分布している(っぽい？)



[Harwath+'17]より抜粋

[P17-1047] Learning word-like units from joint audio-visual analysis

所感

- 人間の子どもが読み書きの習得より先に概念の習得をするのと同じようことをニューラルネットでもできることを示した興味深い研究
  - アライン時にテキスト情報を全く使っていないのに概念に対応するregion-intervalが取れるのが面白い
- このデータセットを公開したらもっと活発になるはず

[P17-1047] Learning word-like units from joint audio-visual analysis


Vancouver Harbour (midnight)

FOIL it! Find One mismatch between Image and Language caption

R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot,
M. Nabi, E. Sangineto and R. Bernardi

マルチモーダルのモデルって本当に複数のモダリティをうまく学習できてるの？

- 入力はそれぞれのモダリティのembeddingの単なるconcatenation
- モデルが何を学習してるかわからない
  - なんなら**言語側だけを使っても**VQA(Visual Question Answering)を解ける
- Image Captioning (IC)の評価指標もn-gramを見てるだけで本質的なcaptionの良さを測っているわけでなはい

やりたいこと：

今のVQAやICのモデルが本当に**2つのモダリティの情報を十分にエンコードできているか検証するためのデータセット＆タスクを設計**

[P17-1024] FOIL it! Find One mismatch between Image and Language caption

FOIL-COCOデータセット

- 正しいキャプション内の1単語を**意味的に似た**他の単語に変えてfoil captionを生成
  - 意味的に似た=objectのsuper-categoryが同じ
  - 例：bird::dog (super-categoryはanimal)

[Shekhar+'17]より抜粋

[P17-1024] FOIL it! Find One mismatch between Image and Language caption

実験

easy

difficult

- T1：キャプションが正しいかを見分ける
- T2：foil captionのfoil wordを当てる
- T3：foil wordを正しいwordに訂正する

人間にはとても簡単なタスクなのに機械は全然ダメ

T1: Classification task			
	Overall	Correct	Foil
Blind ← 言語のみ	55.62	86.20	25.04
CNN+LSTM	61.07	89.16	32.98
IC-Wang IC model	42.21	38.98	45.44
LSTM + norm I	63.26	**92.02**	34.51
HieCoAtt VQAのSOTA	**64.14**	91.89	**36.38**
Human (*majority*)	92.89	91.24	94.52
Human (*unanimity*)	76.32	73.73	78.90

T2: Foil word detection task		
	nouns	all content words
Chance	23.25	15.87
IC-Wang	27.59	23.32
LSTM + norm I	26.32	24.25
HieCoAtt	**38.79**	**33.69**
Human (*majority*)		97.00
Human (*unanimity*)		73.60

T3: Foil word correction task	
	all target words
Chance	1.38
IC-Wang	**22.16**
LSTM + norm I	4.7
HieCoAtt	4.21

[Shekhar+'17]より抜粋

[P17-1024] FOIL it! Find One mismatch between Image and Language caption

所感

- End-to-Endの手法では本質的な問題解決にはなっていないことを明らかにした研究
  - このような研究はとても重要(だと思う)
- Layer-wise Relevance Propagation (LRP)[Bach+'15]を適用して画像とキャプションそれぞれのどこを注目しているかを見てみると面白そう
  - LRPとは，NNの入力層の出力層に対する関連度(寄与度)を求める手法

出力層のノード(クラス)

[Zhang+'16]より抜粋

[P17-1024] FOIL it! Find One mismatch between Image and Language caption

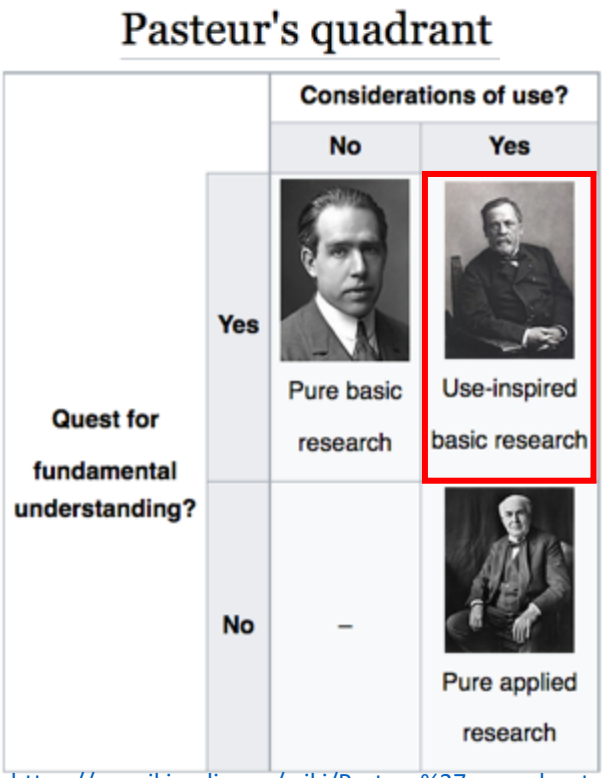
On the Challenges of Translating NLP Research into Commercial Products

Daniel Dahlmeier

A-Maze-ing Laughter

研究を商用化する時の問題点および研究の商用化を増やすための提案

- NLPの研究を商用化することを過小評価している/問題がわかっていない研究者が多い
- どうしたら商用化が増えるのかを著者の経験に基づいて提案
  - **Pasteur’s quadrantの Use-inspired basic research のような研究を目指す**



https://en.wikipedia.org/wiki/Pasteur%27s_quadrant

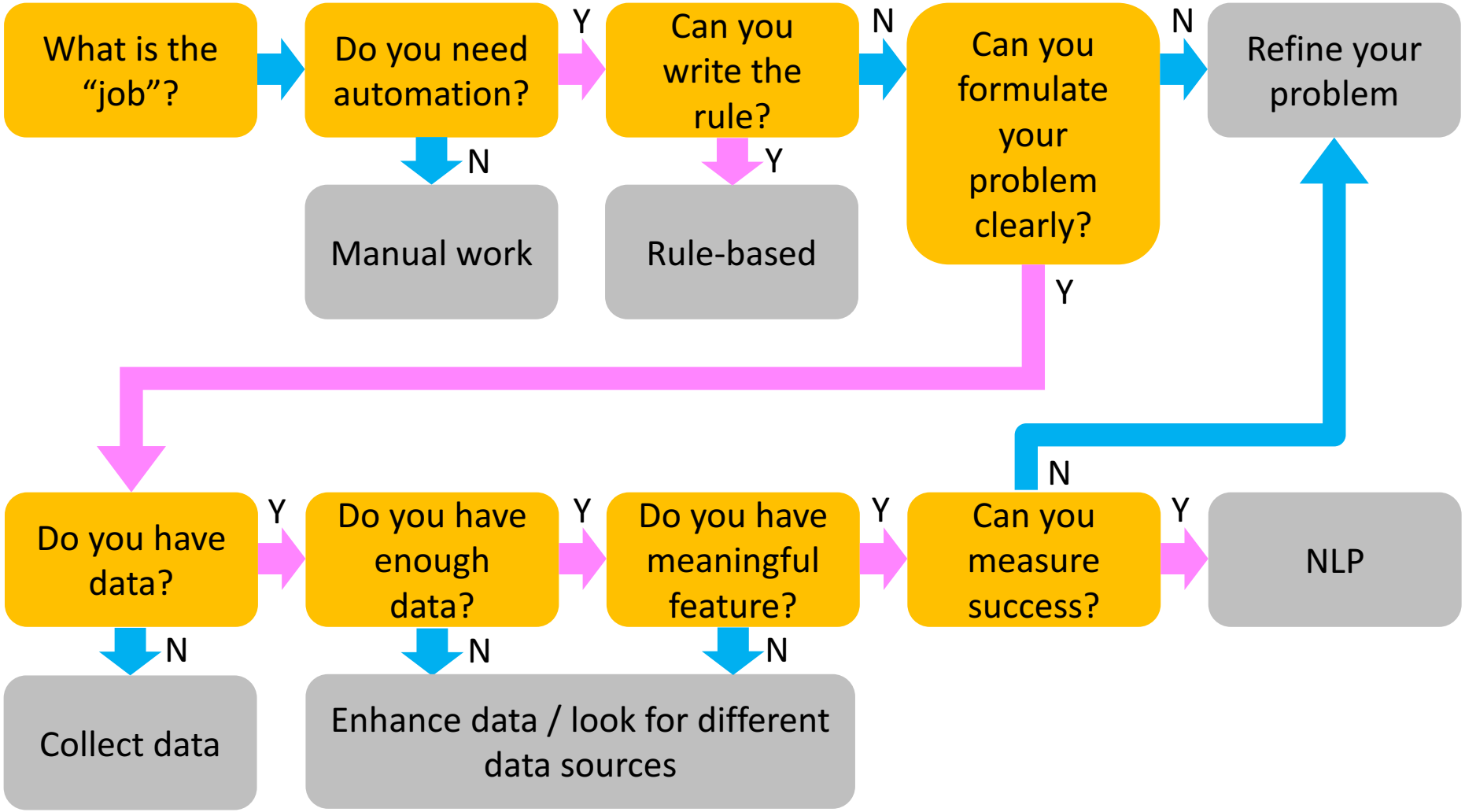
[P17-2015] On the Challenges of Translating NLP Research into Commercial Products

研究を商用化する際に直面する問題点

1. Lack of Value Focus
2. Lack of Reproducibility
3. Lack of (Domain) Data
4. Overemphasis on Test Scores
5. Difficulty of Adoption
6. Timelines

[P17-2015] On the Challenges of Translating NLP Research into Commercial Products

A “Recipe” for Qualifying a Research Problem



[Dahlmeier+’17]の発表スライドをもとに作成

[P17-2015] On the Challenges of Translating NLP Research into Commercial Products

所感

- 新しいアルゴリズムもなければ，実験結果もない斬新な論文．だがとても重要なテーマ
- どこの企業研究者でも同じ問題に直面してるんだなあ

[P17-2015] On the Challenges of Translating NLP Research into Commercial Products

まとめ

- マルチモーダルの研究が増えてきている
  - 新しい手法だけでなく，新しいタスクも
- 本質的な問題を解くタスクの設計が重要
  - 同時にニューラルネットモデルの説明可能な分析手法も考えることが重要(今や強力な武器の一つなので)
- 商用化も見据えた研究(問題)の設計を目指したい

ACLに参加してみた感想


金山さんより提供

- 初めてACL会議に参加
- 人の多さに圧倒された
  - 聴講も含めて1800人くらい
- 毎日coffee break(スタバのコーヒー)*2~3回あったのが良かった
- ポスターセッションが3時間強とか長すぎでは？
  - それでも聞きたい発表を全部回れなかった
- バンクーバーの海辺のサイクリングがとても気持ち良かった
- 来年はメルボルン開催です

ありがとうございました

Lion Gate Bridge

参考文献

- Bach et. al, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”, PLoS ONE, 2015.
- Dahlmeier. “On the challenges of translating NLP research into commercial products”, ACL, 2017.
- Harwath and Glass, “Learning word-like units from joint audio-visual analysis”, ACL, 2017.
- Liu et. al., “Learning character-level compositionality with visual features”, ACL, 2017.
- Mikolov et. al., “Efficient estimation of word representations in vector space”, arXiv, 2013.
- Shekhar et. al., “Foil it! find one mismatch between image and language caption”, ACL, 2017.
- Zhang et. al., “Top-down neural attention by excitation backprop”, ECCV, 2016.

予備スライド

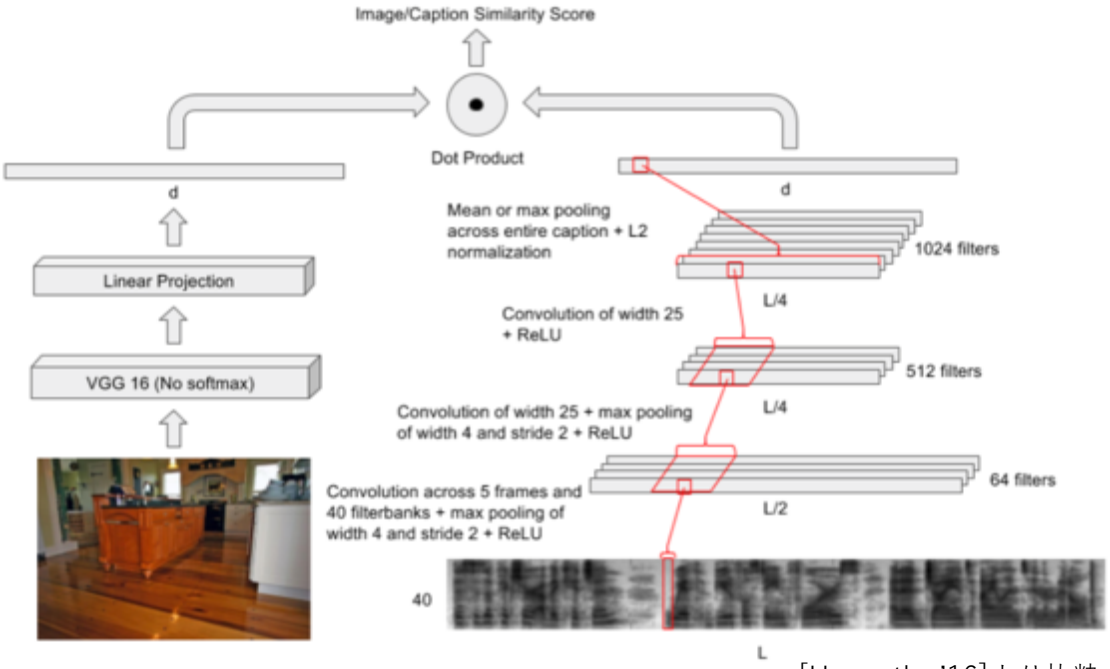
visual-audio model[Harwath+'16]

- 目的関数は負例とのスコアのマージン最大化

captionを置換　　imageを置換

$$\mathcal{L}(\theta) = \sum_{j=1}^{B} \max(0, S_j^c - S_j^p + 1) + \max(0, S_j^i - S_j^p + 1)$$

正例のスコア　　正例のスコア



[Harwath+'16]より抜粋

[P17-1047] Learning word-like units from joint audio-visual analysis

Grounding pairの求め方

- 画像は10x10ピクセルのgridを作り，そこから複数のbboxを選んで一つの候補を作る
- 音声は50/100 framesから選ぶ
- それぞれのモダリティの候補をembedding化して全てのペアの類似度を求める
- 候補の数を減らすため，ノイズ除去と重複した部分をIOUでマージ(音声)
- 画像と音声のembeddingそれぞれ別々にk-meansでクラスタリング
- 式(2)に基づいて，音声のクラスタリングと画像のクラスタリングを結びつける

$$\text{Affinity}(\mathcal{I}, \mathcal{A}) = \sum_{\mathbf{i} \in \mathcal{I}} \sum_{\mathbf{a} \in \mathcal{A}} \mathbf{i}^{\top} \mathbf{a} \cdot \text{Pair}(\mathbf{i}, \mathbf{a}) \quad (2)$$

[P17-1047] Learning word-like units from joint audio-visual analysis

1. Lack of Value Focus

- 良い製品を作る第一歩は自分のカスタマーをよく理解すること
- その研究が可能にするのは何か
- (潜在的な)ユーザにとっての価値は何か

[P17-2015] On the Challenges of Translating NLP Research into Commercial Products

3. Lack of (Domain) Data

- 我々にとってデータは燃料
- なのに．．．
  - お客様データは機密やプライバシーのため公開できない
  - 公開されてるデータを使う際にライセンスや著作権の確認は必須
  - 商用利用不可のデータから学習されたモデルも使えない
  - 実際のデータは少量かつノイズが多く不完全なため，それを使っていい結果を得るのは難しい
- いい結果を出すためには，データのスキーマや，その背後にあるビジネスプロセスを理解しなければならない

[P17-2015] On the Challenges of Translating NLP Research into Commercial Products

4. Overemphasis on Test Scores

- テストデータ上でのスコアは単なるone of many
- その他の考慮すべき基準
  - 実装時間やコスト
  - 計算資源
  - スピードやパフォーマンス
  - 統合容易性
  - 多言語サポート
  - カスタマイズ可能か
  - 事前知識を組み込むことは可能か
  - モデルの解釈・説明は可能か

[P17-2015] On the Challenges of Translating NLP Research into Commercial Products