EMNLP-IJCNLP2019
参加報告

東京大学　鷲尾光樹

自己紹介

- 名前：鷲尾光樹
- 所属：東京大学大学院総合文化研究科言語情報科学専攻
- 学年：D3
- 専門：単語に関する知識獲得（特に意味関係）

- 趣味：漫画執筆
- 多言語漫画配信サービスMantraで英語・中国語で読めます！
  - https://pf.mntr.jp/books/tojime_no_siora
  - https://pf.mntr.jp/books/rasetugari
  - https://pf.mntr.jp/books/balloon_dream

参加報告

- EMNLP-IJCNLP2019@香港について

- 準Best Paper：*Designing and Interpreting Proves with Control Tasks*
  - John Hewitt & Percy Liang

- 閑話休題：香港ディズニーランド

- Best Paper: *Specializing Word Embeddings (for Parsing) by Information Bottleneck*
  - Xiang Lisa Li & Jason Eisner

EMNLP-IJCNLP2019@香港

- 二つのBest PaperはあるタスクのSOTA達成を目指すものではない。
  - Contextualized Representation（ELMo, BERT等）に関わるものではある。

- 準Best：
  - Probing taskにおけるAccuracy比較の妥当性について

- Best：
  - Contextualized Representationから、対象タスクに関連する圧縮表現を抽出する方法

Designing and Interpreting Probes with Control Tasks

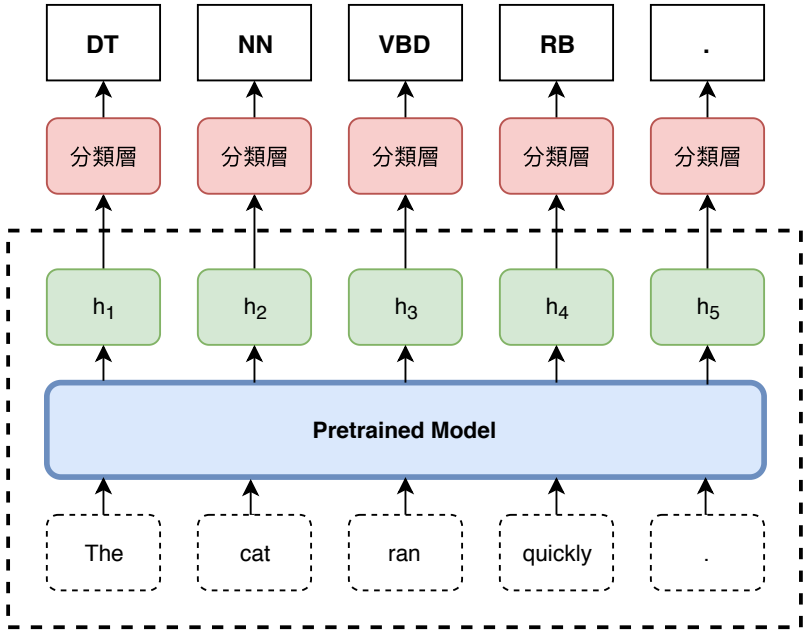
John Hewitt
Stanford University
johnhew@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

- Probingを単純な性能比較でやってしまって大丈夫か？

- Probing
  - 動機：ある表現（特にELMo、BERT）が（統語的・意味的）文法知識を符号化できているか確かめたい。

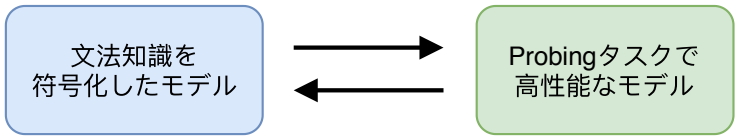
  - 方法：表現をfixして、分類層のみを教師あり学習し、性能を見る。



点線内はFix
（教師あり学習時に更新しない）

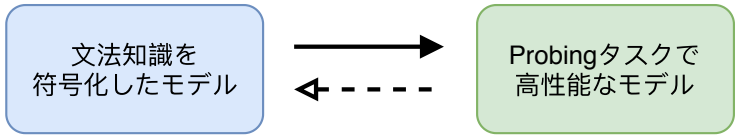
Probingの前提と妥当性

- 文法知識を符号化したモデル＝Probingタスクで高性能なモデル

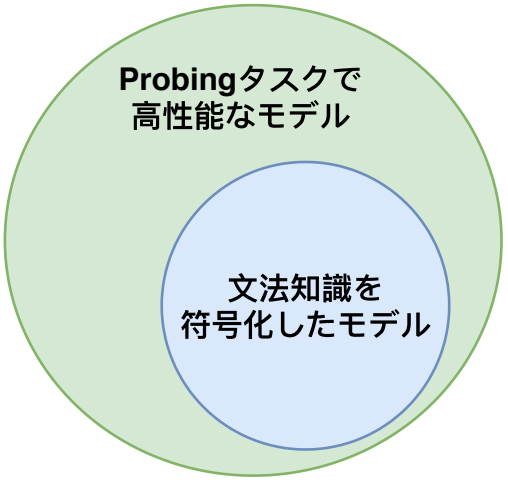


- しかし、文法知識を符号化していなくても高性能なモデルが存在するのでは？
  - Pretrainedモデルの埋め込みが入力の情報を十分に保持しており、
  - 分類層が十分に強力で、
  - データ量が十分ならば？

こっちの条件は大丈夫そう



こちらの条件は成り立つか？

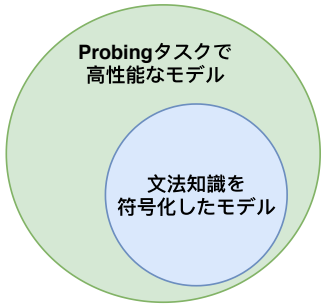


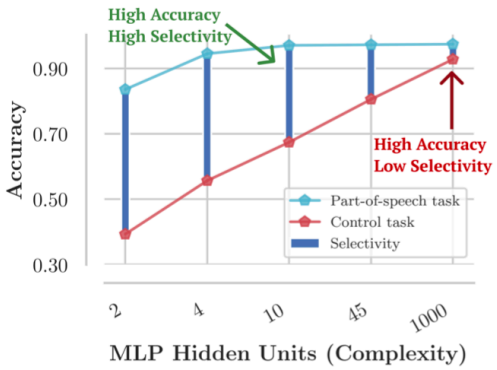
ProbingのSelectivityとControlタスク

- あるProbingタスクに対応する**Controlタスク**：
  - 出力（ラベル）を対象の文法知識とは独立かつ恣意的なものに置換したタスク

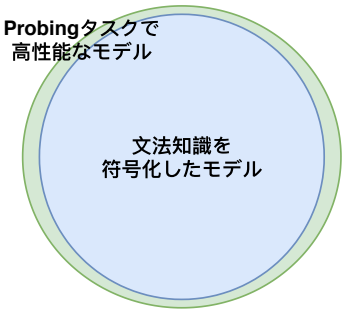
- あるProbing（データセット、分類層の選択、最適化・正則化法等の設定を含める）の**Selectivity**：
  - ProbingタスクのAccuracyとControlタスクのAccuracyの差

  - high linguistic acc & high control acc
    - 分類層が強力でデータ量も多い等で、
    - 文法知識とは独立のControlタスクも
    - 解けてしまう場合。





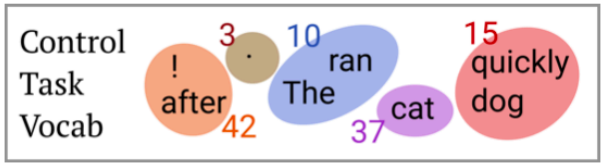
  - **high linguistic acc & low control acc**
    - **分類層が貧弱、データ量が少ない等で、**
    - **符号化された文法知識を用いてしか**
    - **解けない場合**



Controlタスクの構築法（POS taggingのみ説明）

- 「構造」を持ったタスクにすること
  - 出力（ラベル）はWord identityによって決まるようにする。
- ランダム性を持ったタスクにすること（恣意性の担保）
  - ある単語のラベルはランダムに決める。
  - ラベル空間はもとのProbingタスクに揃える。
    - POS tagラベル数が45個なので、45個のラベルからランダムサンプリング
- 入力事例とラベルの数は同じだが、統語知識ではなく単語のみで分類可能なタスクができる。

※Controlタスクでは、
訓練データに出てきた単語のみ
しかうまく分類できない。



Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	10	37	10	15	3
Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	10	15	10	42	42

実験（ELMoの一層目で実験）

- タスク：POS tagging(PoS)＆Dependency edge prediction(Dep)

- デフォルトのProbing（すべての訓練データを使い、正則化はなし）は良いProbingか？
  - high linguistic accだが…
    - 分類層にMLPを用いると低Selectivity
  - Depでは、Linearに比べてMLPが対象タスクで高性能だが、Selectivityはガタ落ち
    - 性能向上はELMoの文法知識由来ではなさそう

Probe	PoS	Ctl	Select.	Dep	Ctl	Select.
Probes with Default Hyperparameters						
Linear	97.2	71.2	26.0	-	-	-
Bilinear	-	-	-	89.0	82.4	6.6
MLP-1	97.3	92.8	4.5	92.3	93.0	-0.7
MLP-2	97.3	93.2	4.2	93.9	92.0	1.9

- タスクの設定をいじって、良いProbing（high linguistic acc & high selectivity）にできるか？
  - 可能
    - 訓練データ量減
    - 分類層のランク・隠れ状態サイズ減など

Probes Designed with Control Tasks						
Linear	97.0	64.0	33.0	-	-	-
Bilinear	-	-	-	91.0	83.1	7.9
MLP-1	97.2	80.6	16.6	90.5	84.3	6.2
MLP-2	97.2	81.7	15.4	92.8	89.8	3.0

実験（ELMo一層目と二層目を比較）

- 背景：Probingでは、ELMo一層目(ELMo1)の方がPoSに関する統語知識を捉えていると言われている。

- ELMo1の方がわずかに対象タスク性能が高い
- 一方で、ELMo2はSelectivityにおいて大きく上回っている。

	Part-of-speech Tagging			
	Linear		MLP-1	
Model	Accuracy	Selectivity	Accuracy	Selectivity
Proj0	96.3	20.6	97.1	1.6
ELMo1	97.2	26.0	97.3	4.5
ELMo2	96.6	31.4	97.0	8.8

- ELMo1の方が統語知識をよく捉えられているとは言えないのでは？
  - 単純にどの単語がどのPOSになりやすいかを覚えることでわずかに性能が向上している可能性がある。

- Linguistic taskのAccのみで、Probingしない方がよい。

まとめと感想

- Probingに対してSelectivityという新たな分析軸を提供
  - POS taggingとDep edge predictionでControlタスクの構築法を開発
  - 分析例を提示

- 感想
  - 第一印象としては含意関係認識や機械読解でやられてる分析をPoSやDepのProbingでもやってみましたという感じ。
    - BERT等の性能が良いからといってそのタスクが解けているのか？
      - 機械読解には簡単な事例が多い[Sugawara+, 2018]
      - 含意関係認識を単語の重複等を見て解いてしまっている[McCoy+, 2019]
    - ※ただ、これらはPretrainedモデル内の知識を見ようとする話ではない。
  - **出力（ラベル）をいじる枠組み（Controlタスク）と**Selectivityという軸の提案は重要。
  - ただ、Selectivityの運用は難しそう。
    - Selectivityと対象タスクの性能がどのくらい良ければGood Probingなのか
    - どのくらいSelectivityに差があれば、タスク性能差由来の結論を棄却できるのか

閑話休題：香港ディズニーランド

- 会場から車で15分くらい。
- 16時から入れて23時まで遊んでよい。
- 夜はトゥモローランドを貸し切り。
  - 食事とお酒が出る。
- MARVEL＆STAR WARS好きの方におすすめ！
  - アトラクションが豊富

閑話休題：香港ディズニーランド

Specializing Word Embeddings (for Parsing) by Information Bottleneck

Xiang Lisa Li
Department of Computer Science
Johns Hopkins University
xli150@jhu.edu

Jason Eisner
Department of Computer Science
Johns Hopkins University
jason@cs.jhu.edu

- Contextualized Representationからあるタスク（論文ではParsing）に関連する圧縮表現を、Information Bottleneck法で抽出

- **Information Bottleneck法**
  - 確率変数X,Yについて、Yを復元できるようにXを圧縮する。
  - 圧縮表現TはXの中で、Yに関連する情報を持つ部分となる。
  - 以下を最小化するようにXをTに圧縮する：

$$\mathcal{L}_{IB} = -\mathrm{I}(Y;T) + \beta \cdot \mathrm{I}(X;T)$$

    - ただし、I(X, Y)はXとYの相互情報量

Information Bottleneck法（IB法）

- Xを国際会議、Yを国際会議の満足度とする。
  - 機械翻訳の研究に熱心なAさんの満足度: $Y_A \in [0, 1]$
  - 不真面目なBさんの満足度: $Y_B \in [0, 1]$

$X = \{$機械翻訳の研究数, 要約の研究数, 意味解析の研究数, …,
料理のおいしさ, バンケットがディズニーかどうか$\}$

- このとき、AさんとBさんの満足度をそれぞれ予測するために、どのようにXを圧縮できるか？

Information Bottleneck法（IB法）

X = {機械翻訳の研究数, 要約の研究数, 意味解析の研究数, …,
料理のおいしさ, バンケットがディズニーかどうか}

- Y_Aを予測するためにXの関連する部分をまとめたT_A:

- Y_Bを予測するためにXの関連する部分をまとめたT_B:

Information Bottleneck法（IB法）

X = {機械翻訳の研究数, 要約の研究数, 意味解析の研究数, …,
料理のおいしさ, バンケットがディズニーかどうか}

- Y_Aを予測するためにXの関連する部分をまとめたT_A:

T_A = {機械翻訳の研究数, $Seq2Seq$を用いた研究数,
$BERT$を用いた研究数, …}

- Y_Bを予測するためにXの関連する部分をまとめたT_B:

Information Bottleneck法（IB法）

X = {機械翻訳の研究数, 要約の研究数, 意味解析の研究数, …,
料理のおいしさ, バンケットがディズニーかどうか}

- Y_Aを予測するためにXの関連する部分をまとめたT_A:

T_A = {機械翻訳の研究数, $Seq2Seq$を用いた研究数,
$BERT$を用いた研究数, …}

- Y_Bを予測するためにXの関連する部分をまとめたT_B:

T_B = {料理のおいしさ, バンケットがディズニーかどうか}

Information Bottleneck法（IB法）

- 最小化する関数：

$$L_{IB} = -I(Y;T) + \beta \cdot I(X;T)$$

大きくすることで
圧縮表現TとYの依存させる
（TとYを関連させる）

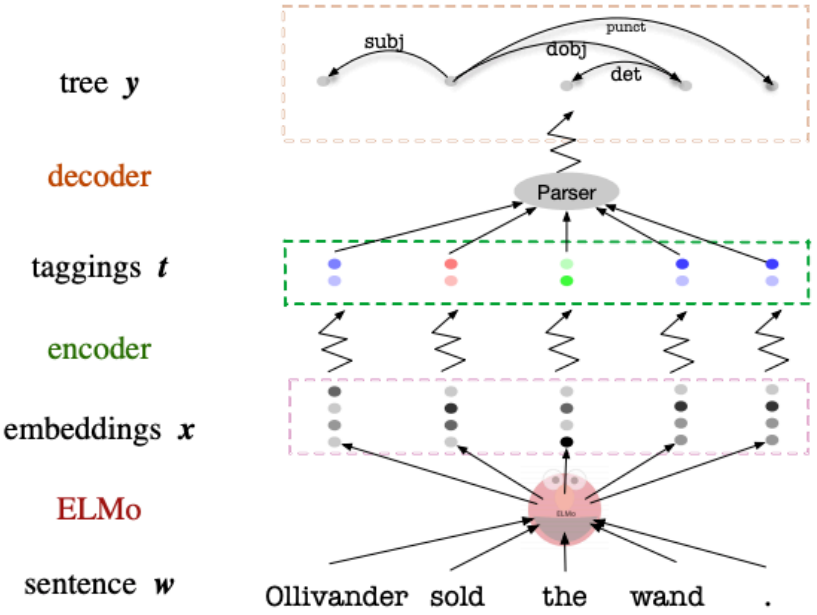
小さくすることで
圧縮表現TとXの独立にする
（情報を落として圧縮する）

  - βは二項のトレードオフを制御するハイパーパラメタ
    - 小さすぎるとI(Y;T)はI(Y;X)に近づく。（圧縮されない）
    - 大きすぎるとTでYを予測できなくなる。（圧縮されすぎ）

- 深層学習の場合はVariational IB法（VIB法）が使える[Alemi+, 2016]

ELMo、依存構造解析、IB法

- 本研究では、
  - X: ELMoから得られる埋め込み
  - Y: 依存構造木
  - T: 連続(埋め込み)or離散(one-hot)な圧縮表現



- Tは依存構造解析に関わる統語的情報のみを捉えた表現になるはず。
  - ELMo内の意味的情報は圧縮時に落ちるはず。

Encoder：X -> T

- ELMo埋め込みを入力としたFeedforward networkで$p_\theta(t_i|x_i)$をモデル化
  - 連続の場合：ガウス分布（d次元平均ベクトルと共分散行列の対角成分を出力）
  - 離散の場合：k次元のsoftmax分布（k次元のベクトルを出力）

- XとTの相互情報量

$$I(X;T) = E_x\left[E_{t\sim p_\theta(t|x)}\left[\log \frac{p_\theta(t|x)}{p_\theta(t)}\right]\right]$$

ここの推定が大変

- パラメータを割り当てた変分分布rを導入し、変分上界を導く。

$$\underbrace{\mathbb{E}_x\left[\mathbb{E}_{t\sim p_\theta(t|x)}\left[\log \frac{p_\theta(t|x)}{r_\psi(t)}\right]\right]}_{\text{upper bound}} - \underbrace{\mathbb{E}_x\left[\mathbb{E}_{t\sim p_\theta(t|x)}\left[\log \frac{p_\theta(t\mid x)}{p_\theta(t)}\right]\right]}_{\mathrm{I}(X;T)}$$

$$= \mathbb{E}_x\left[\mathrm{KL}(p_\theta(t) \,||\, r_\psi(t))\right] \geq 0$$

$$\mathbf{KL}(p_\theta(t|x)||r_\psi(t))$$

二つの確率分布は陽にわかるので、
これは簡単に計算できる！

Decoder: T -> Y

- 圧縮表現の系列tを入力とし、依存構造木yを復号化
  - biaffine dependency parser[Dozat&Manning, 2016]
  - 下記の$q_\phi(y|t)$をモデル化

- YとTの相互情報量

$$I(Y;T) = E_{y,t\sim p_\theta}\left[\log \frac{p_\theta(y|t)}{p(y)}\right]$$

これを大きくすればよいが
xの周辺化が関わるので
推定が大変

ここはθと無関係

- $p_\theta(y|t)$を$q_\phi(y|t)$で変分近似し、下界を導く

$$\overbrace{\mathbb{E}_{y,t\sim p_\theta}[\log \frac{p_\theta(y|t)}{p(y)}]}^{\mathrm{I}(Y;T)} - \overbrace{\mathbb{E}_{y,t\sim p_\theta}[\log \frac{q_\phi(y|t)}{p(y)}]}^{\text{lower bound}}$$

$$= \mathbb{E}_{t\sim p_\theta}[\mathrm{KL}(p_\theta(y \mid t) \,||\, q_\phi(y \mid t))] \geq 0$$

$$\Rightarrow \quad \mathbb{E}_{t\sim p_\theta(t|x)}[-\log q_\phi(y|t)]$$

tを$p_\theta(t|x)$から
サンプリングして推定

モデルの訓練と実験

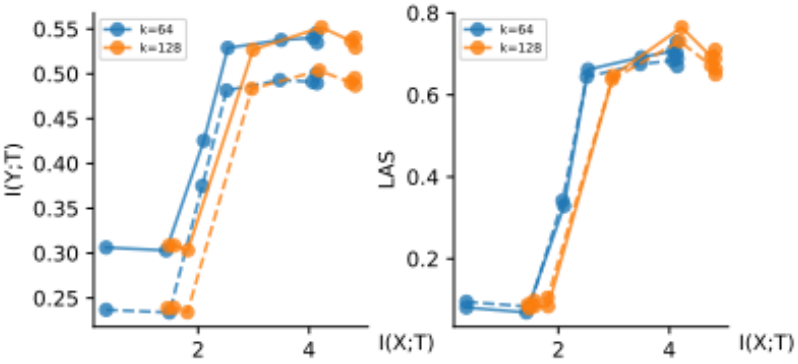
- 最小化する目的関数（※実はもう一つ項があるが省略）

$$L = E_{x,y}\left[E_{t\sim p_\theta(t|x)}\left[-\log q_\phi(y|t)\right] + \beta \cdot KL(p_\theta(t|x)||r_\psi(t))\right]$$

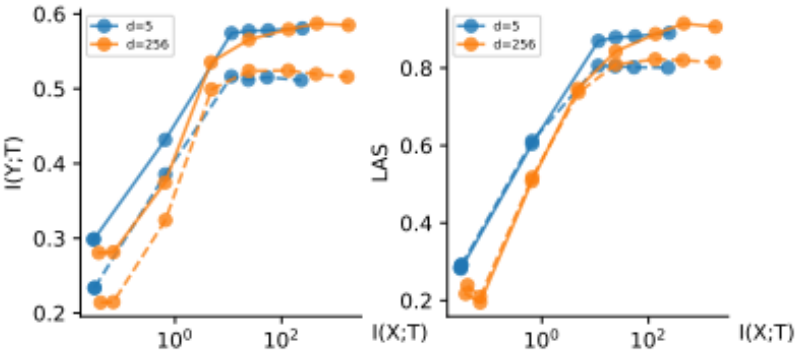
- Universal dependencies内の9言語について訓練
  - ELMoはfix

Scientific Evaluation

- β を動かしたときのトレードオフカーブ
  - I(X;T)を上げていったときにI(Y;T)と性能（LAS）が頭打ちになる点がある。
  - ELMo表現内で依存構造解析に関わる部分には限りがある。
  - 5次元連続値表現でも256次元のものと同等の性能（CPUで高速に動く）



(a) Discrete Version

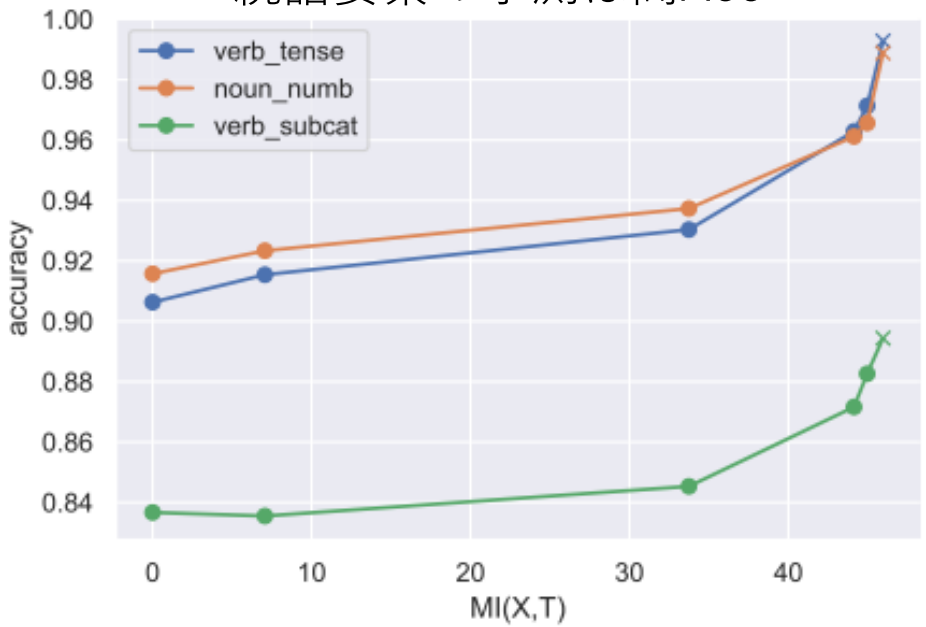


(b) Continuous Version

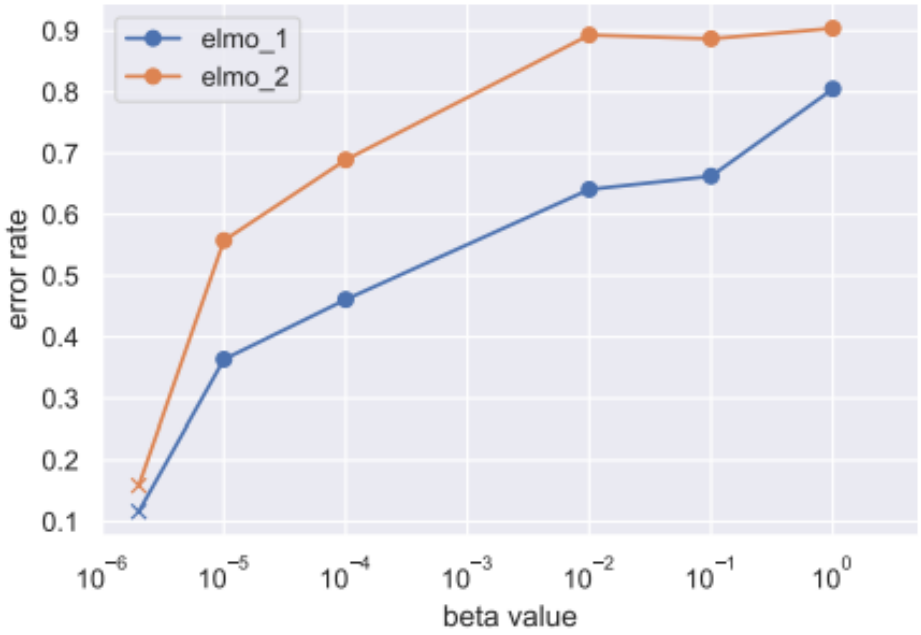
Scientific Evaluation

- Tから時制などは予測できるが、語幹を当てたりはできない。
  - 統語的な情報のみが符号化されており、意味的な情報が適切に落ちている。

統語要素の予測は高Acc



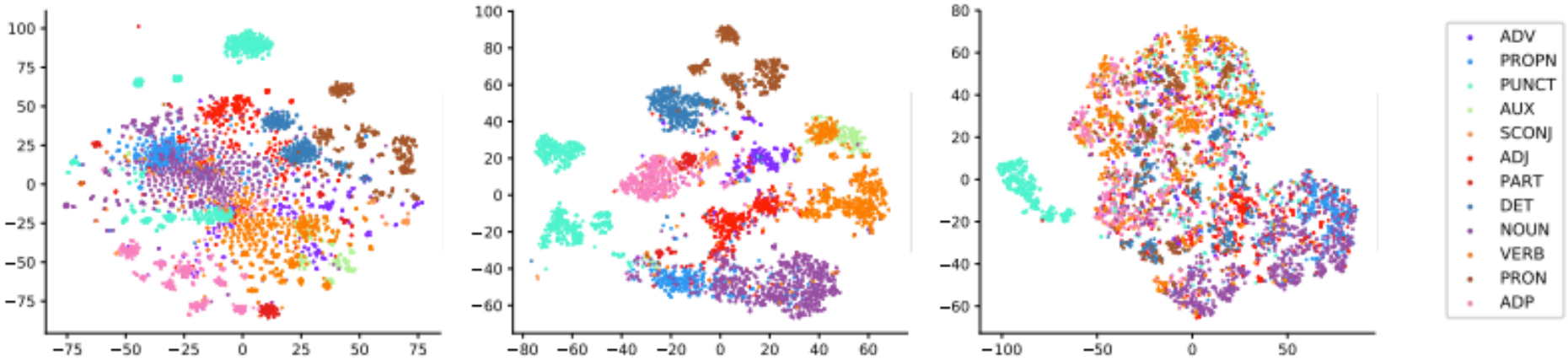
語幹の予測は高Error rate



Scientific Evaluation

- TにはPOSの情報も適切に符号化されている。
  - POSは直接学習していない！

未圧縮　　ほどよい圧縮　　過剰な圧縮



(a) ELMo, $\mathrm{I}(X;T) = H(X) \approx 400.6$　　(b) $\mathrm{I}(X;T) \approx 24.3$　　(c) $\mathrm{I}(X;T) \approx 0.069$

- 未圧縮では統語情報以外も混在

Engineering Evaluation

Models	Arabic	Hindi	English	French	Spanish	Portuguese	Russian	Chinese	Italian
Iden	0.751	**0.870**	0.824	0.784	0.808	0.813	0.783	0.709	**0.863**
PCA	0.743	**0.866**	0.823	0.749	0.802	0.808	0.777	0.697	0.857
MLP	0.759	**0.871**	0.839	0.816	**0.835**	0.821	0.800	0.734	**0.867**
VIBc	**0.779**	**0.866**	**0.851**	**0.828**	**0.837**	**0.836**	**0.814**	**0.754**	**0.867**
POS	0.652	0.713	0.712	0.718	**0.739**	**0.743**	**0.662**	0.510	0.779
VIBd	**0.672**	**0.736**	**0.742**	**0.723**	**0.725**	0.710	**0.651**	**0.591**	**0.781**

- 提案手法：VIB
- Iden（ELMoそのまま）より良い。
- 他の次元削減手法（PCA、MLP）よりも良い。

まとめと感想

- IB法によってContextualized representationから依存構造解析に関わる圧縮表現を抽出
  - 解釈しやすい分析結果
  - 工学的にも良い特徴

- 感想
  - 分析が大量＆捻り出した感がなく綺麗。
  - 論文内でもParsing特有の手法ではないことに言及しており、XとYに色々なものを当てはめると面白そう。
    - 日本語文と英語翻訳文
    - 発話と応答
    - etc...