

Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning

Angeliki Lazaridou, Georgiana Dinu, Marco Baroni
ACL-IJCNLP 2015 (図表・式はこの論文より引用)

発表者: 小町守 (首都大学東京)

[<komachi@tmu.ac.jp>](mailto:komachi@tmu.ac.jp)

第7回最先端 NLP 勉強会@SmartNews

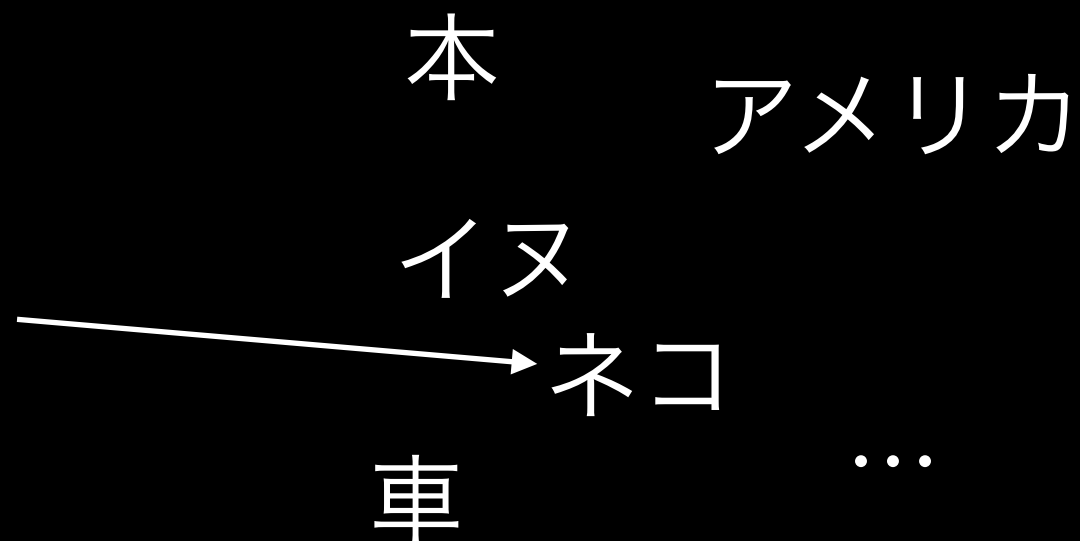
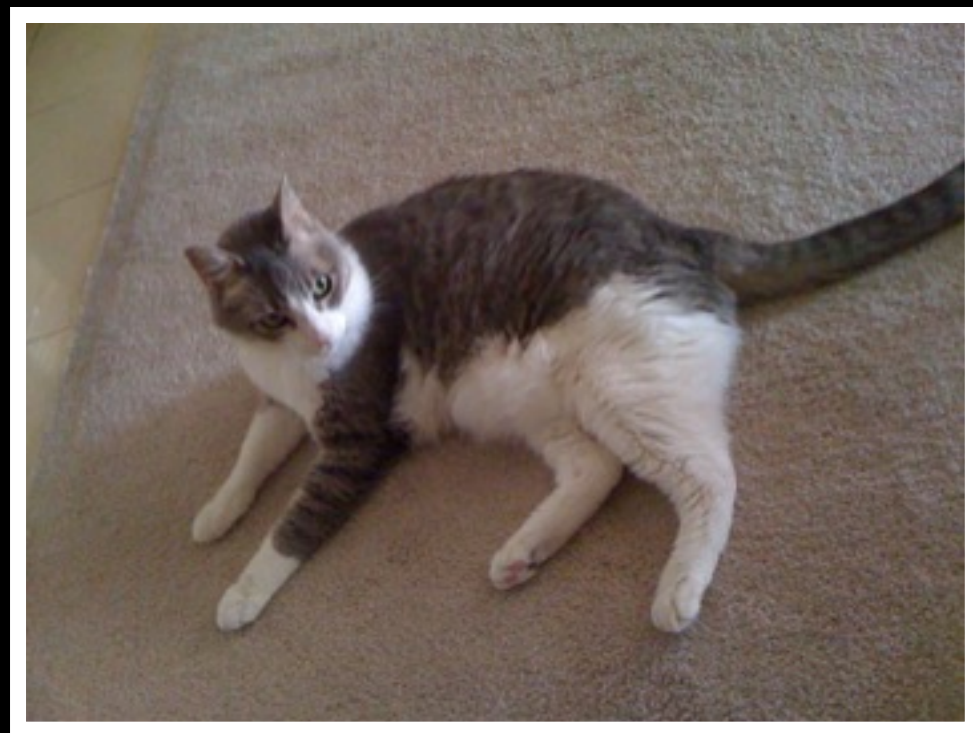
August 29th, 2015

たくさんラベルがある場合 ふつうの教師あり学習は厳しい

- ふつうの教師あり学習

(x, y) : x が素性ベクトル、 y がラベル

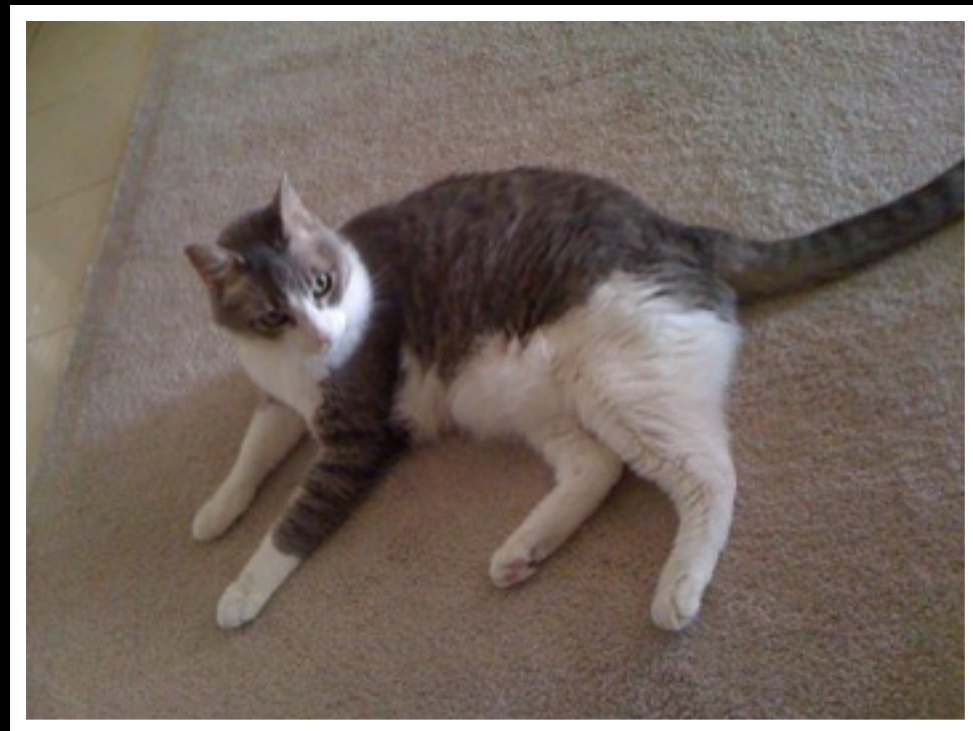
→ラベル間に関係がなく、ラベル集合が巨大な
ときにアノテーションがスケールしない



そこで Zero-shot learning

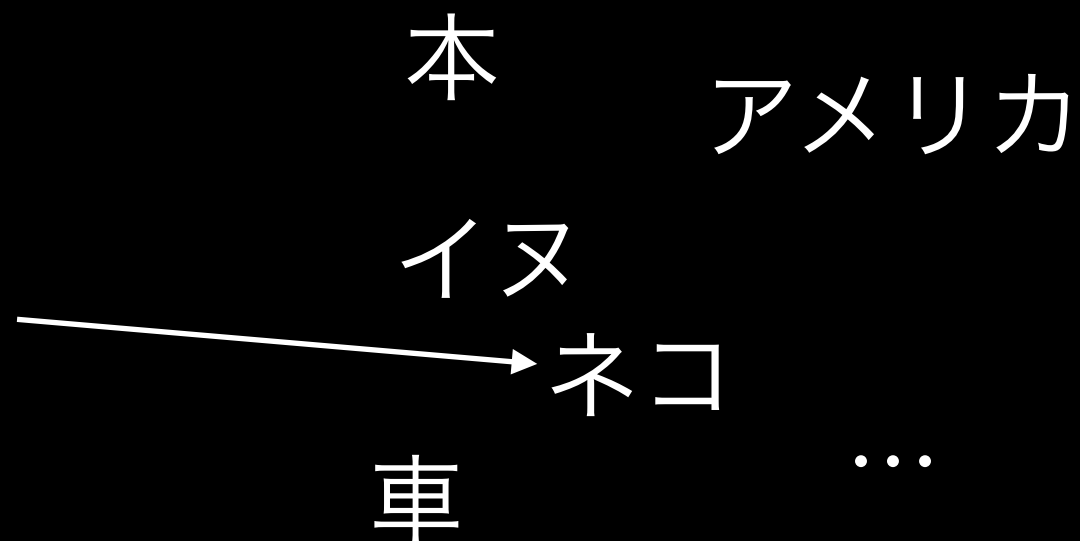
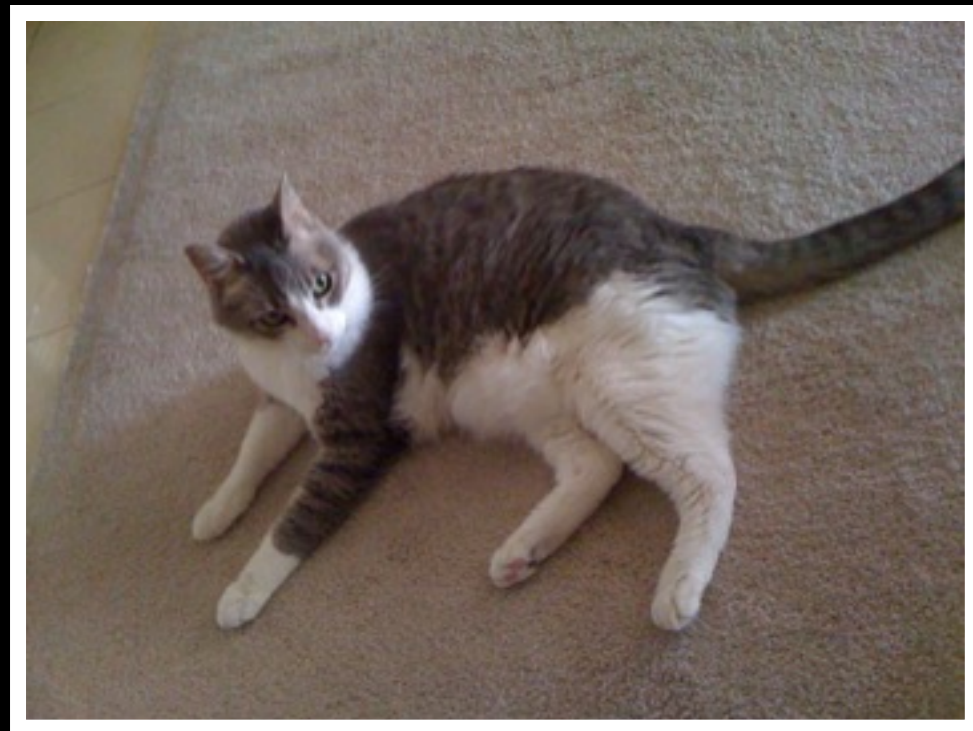
- Zero-shot learning

(x, y) : y は単語（やフレーズ）で、ラベル間に関係があるので、分布類似度を用いてラベルの意味空間を推定することで解決



Zero-shot learning は どのように学習する？

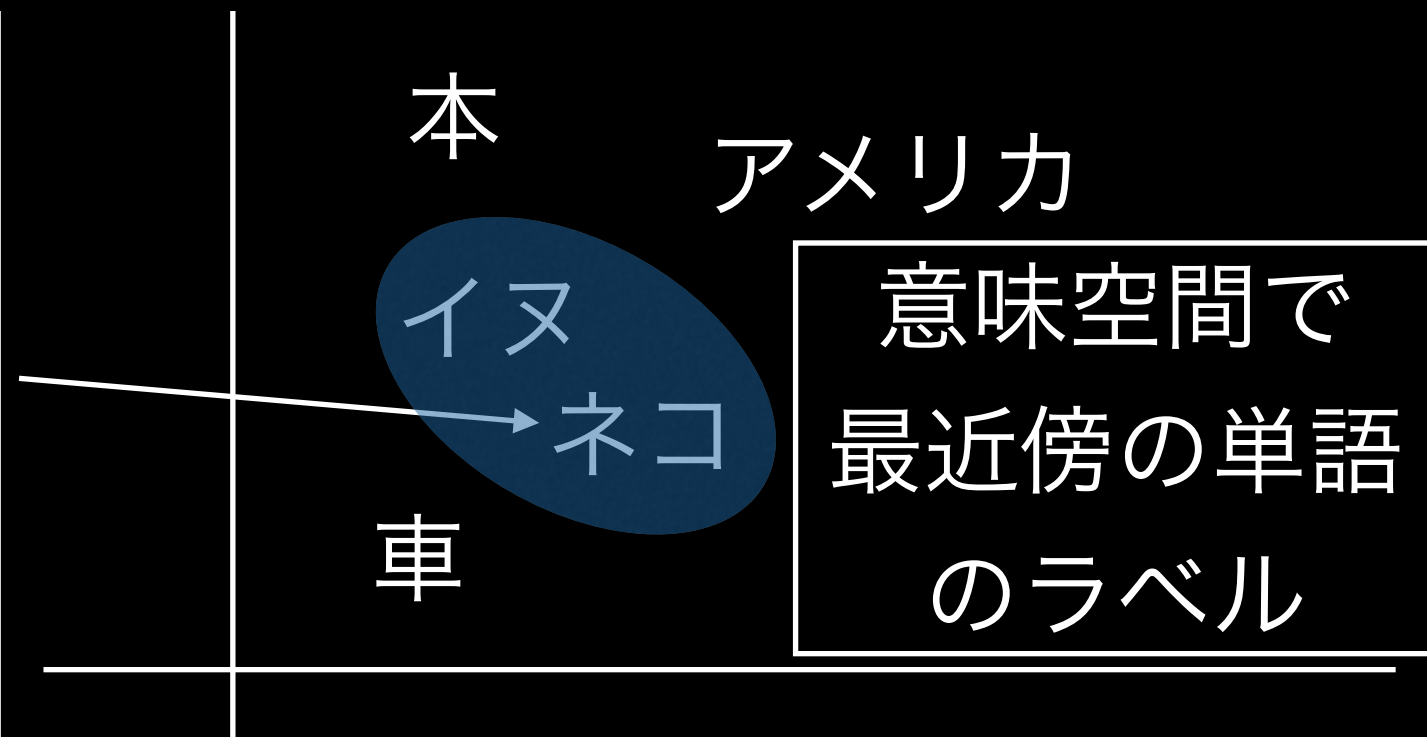
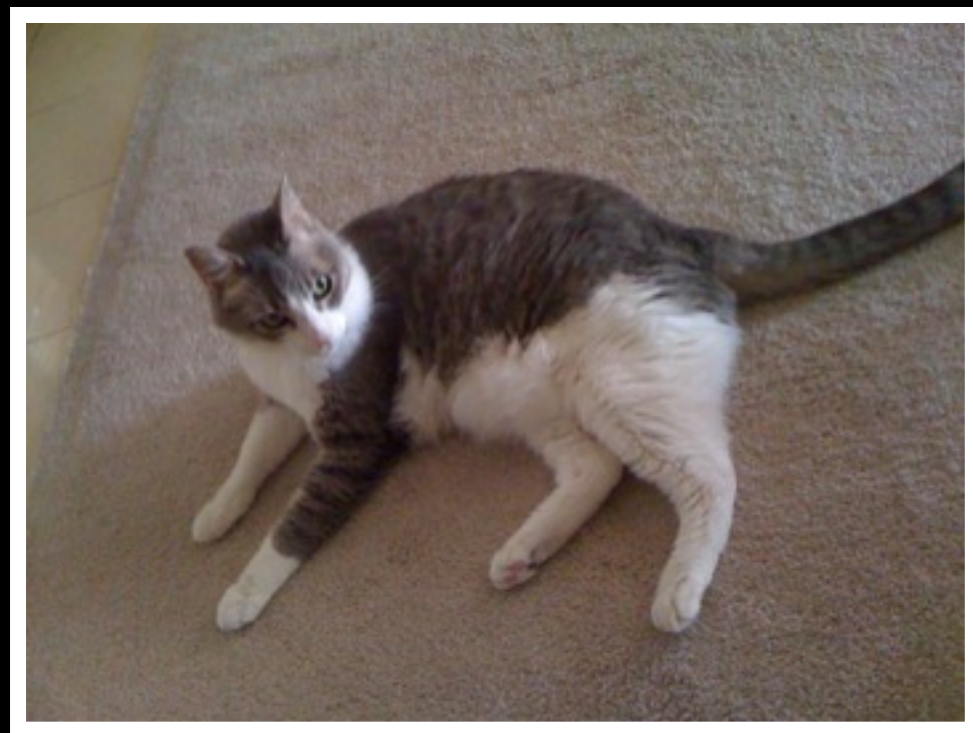
- ふつうの教師あり学習
→ 入力 of 素性空間から相互に無関係なラベルを
予測する関数を求める



Zero-shot learning は どのように学習する？

- Zero-shot learning

→ 入力の素性空間のベクトルから分布類似度で求めた意味空間で対応する単語ラベルにマップする関数 (cross space mapping) を求める



分布類似度を用いた Zero-shot Learning タスク

- ・ 脳信号のデコード (ZSL の最初の研究)
(Mitchell et al., 2008)
- ・ 画像ラベリング (クロスモーダル)
(Frome et al., 2013; Lazaridou et al., 2014;
Socher et al., 2013)
- ・ 対訳辞書・フレーズテーブル構築 (言語横断)
(Dinu and Baroni, 2014; Mikolov et al., 2013)

本研究の主要な貢献

- Zero-shot learning における線形のクロスモーダルマッピング関数の性質について考察
 - 2乗誤差最小法では「ハブ」が出現！
- マージン最大ランキングロス関数で改善



本研究の主要な貢献

- Zero-shot learning における線形のクロスモーダルマッピング関数の性質について考察
 - クロスモーダルでは汚染（過学習）が出現！
- 自己学習によるデータ拡張で改善



ただし負例をうまく
選ぶ必要がある

空間横断マップ学習における ロードマップ

表 1: 本研究の提案手法による精度向上 (Precision@1)

	言語横断	クロスモーダル
state-of-the-art	33.0	0.5
通常のマッピング	29.7	1.1
最大マージン (3節)	39.4	1.9
データ拡張 (4節)	NA	3.7
負例の利用 (5節)	40.2	5.6

Zero-shot learning の 目的関数

- ・ リッジ回帰

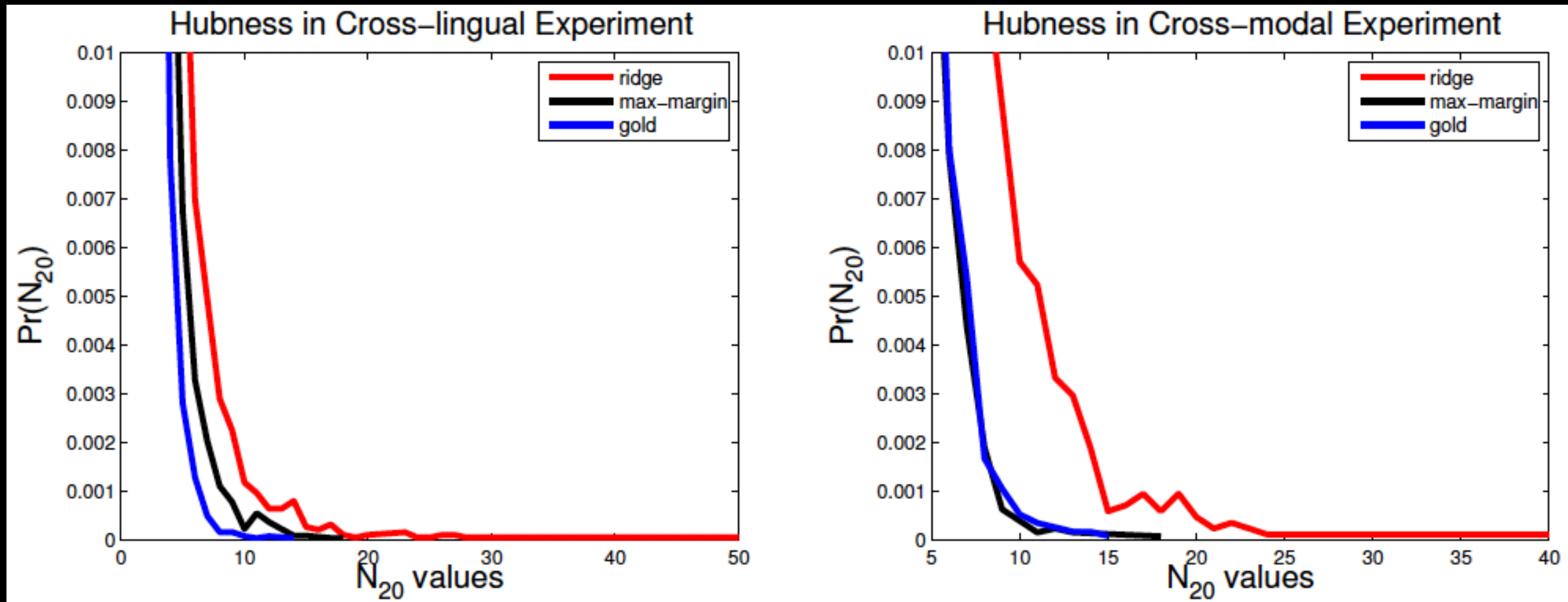
$$\hat{\mathbf{W}} = \operatorname{argmin}_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\| + \lambda \|\mathbf{W}\|$$

- ・ マージン最大化

$$\sum_{j \neq i}^k \max\{0, \gamma + \operatorname{dist}(\hat{\mathbf{y}}_i, \mathbf{y}_i) - \operatorname{dist}(\hat{\mathbf{y}}_i, \mathbf{y}_j)\}$$

ただし dist は距離関数（ここでは \cos 類似度の逆数）
 γ はマージン、 k は負例の数に対応するハイパーパラメータ
→全てのラベルで和を取るのはつらいのでサンプリングする

高次元空間では必ずハブが出現 (Radovanovic et al., 2010)



• ただし $N_k(y) = |\{x \in T | y \in NN_k(x, S)\}|$

なぜリッジ回帰では ハブが出現するのか？

定理 2 リッジ回帰の写像行列を $M \in \mathbb{R}^{d \times c}$, 説明変数の行列 $A \in \mathbb{R}^{c \times n}$, 目的変数の行列 $B \in \mathbb{R}^{d \times n}$ と定義した場合, 写像行列は

$$M = \arg \min_M (\|MA - B\|_F^2 + \lambda \|M\|_F). \quad (4)$$

によって求まる. ここで, $\lambda \geq 0$ は正則化パラメータである. このとき, 写像された説明変数と目的変数の関係は $\|MA\|_2 \leq \|B\|_2$ である.

本稿では, データが中心化されていることを想定しているので, 行列の 2-ノルムは主成分方向の分散を示す尺度として解釈できる. 従って, 定理 2 は, 写像された説明変数 MA の主成分方向の分散が, 目的変数 B の分散よりも小さくなりやすいことを示している.

さらに, 正則化項が無い場合 ($\lambda = 0$) においても, $\|MA\|_2 \leq \|B\|_2$ は成り立ち, 射影された説明変数の分散が小さくなりやすいという傾向が生じる. その結果, 単純に正則化パラメータ $\lambda = 0$ としても, 写像された説明変数の分散が小さくなるという傾向を完全に排除することはできない.

既存法は, A を訓練事例集合の行列 $X = [x_1 \cdots x_n] \in \mathbb{R}^{c \times n}$, B を訓練ラベル行列 $B = Y = [y_1 \cdots y_n] \in \mathbb{R}^{d \times n}$ として, A を (写像行列 M によって) ラベル空間へ写像する. 上述した定理 2 より, A は B よりも原点に近い位置に写像される. 定理 2 は訓練セットのみについて議論しているものの, 写像された (X に含まれない) 評価事例も多くの (Y に含まれる) 訓練ラベルよりも, 原点に近い位置に写像される傾向にあることを示唆している.

線形回帰 (リッジ回帰) だと、マップした空間の原点に近い位置に写像されやすくなってしまうため (重藤ら, NL研2015)

最大マージン法によって ハブの問題は軽減できる

表 3: リッジ回帰と最大マージン法の比較

	言語横断		クロスモーダル	
	リッジ	最大マージン	リッジ	最大マージン
P@1	29.7	38.4	1.1	1.9
P@5	44.2	54.2	4.8	5.4
P@10	49.1	60.4	7.9	9.0

表 4: テスト事例が最近傍となっている割合

	言語横断			クロスモーダル		
	リッジ	最大マージン	ゴールド	リッジ	最大マージン	ゴールド
	19.6	9.8	0.6	55.8	21.6	7.8




クロスモーダル環境だと、 ハブだけではなく汚染も問題

	ラベル数20万		ラベル数5,000
	言語横断	クロスモーダル	クロスモーダル
汚染度	8.7%	18%	88%

- ・ 「汚染」：訓練事例のラベルをテスト事例につけてしまう（kNNの中に入れてしまう）こと
- ・ 探索空間を小さくしようとする、汚染がさらに深刻になる

画像キメラを用いた自己学習

- ・ 訓練事例にない y_i に対し、近傍の訓練事例で出現した単語ベクトルに対応する画像ベクトルの平均を用いて、画像ベクトル x_i を推定
- ・ 深層学習でも「浅い」学習でも効果があり、最近では NLP でも適用されている (Zhang and LeCun, 2015)

dolphin	tarantula	highland
		
whale orca porpoise cetacean shark	anteater arachnid spider opossum scorpion	whisky lowland bagpipe glen distillery

言語の類似性と画像の類似性は異なるので、ノイジーなデータが生成される (上図の opossum や bagpipe)

画像キメラによってクロス モーダル zero-shot learning が改善

表 7: 最大マージン法と画像キメラによるクロスモーダル実験

	なし	キメラ5	キメラ10
P@1	1.9	3.7	3.2
P@5	5.4	10.9	10.5
P@10	9.0	15.8	15.9

	なし	キメラ5	キメラ10
汚染度	88%	71%	73%

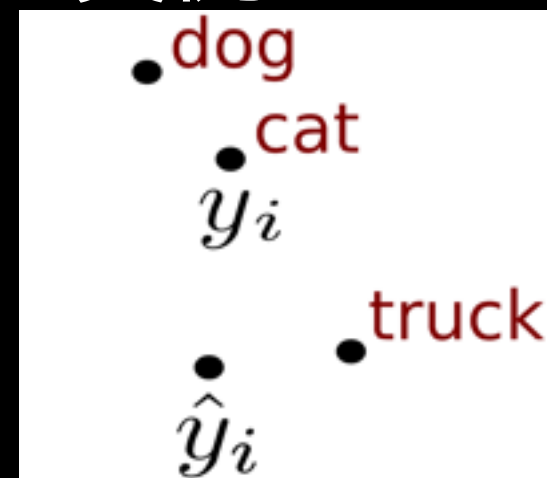
- ・ クロスモーダルタスクでは汚染度が下がり適合率も向上
- ・ もともと汚染度が低い言語横断タスクでは逆に適合率低下 (表10=割愛)

最大マージン法では 負例の選び方が性能に影響する

- ・ ロス最小化 (Crammer et al., 2006)
→ もっとも性能に影響を与えそうな事例を負例として選択する

- ・ $s_j = \cos(\hat{y}_i, y_j) - \cos(y_i, y_j)$
がもっとも高い s_j をもつ事例を負例に用いる

- ・ 図2 (右図) において、cat が正解だとすると、
dog ではなく truck を負例に用いる



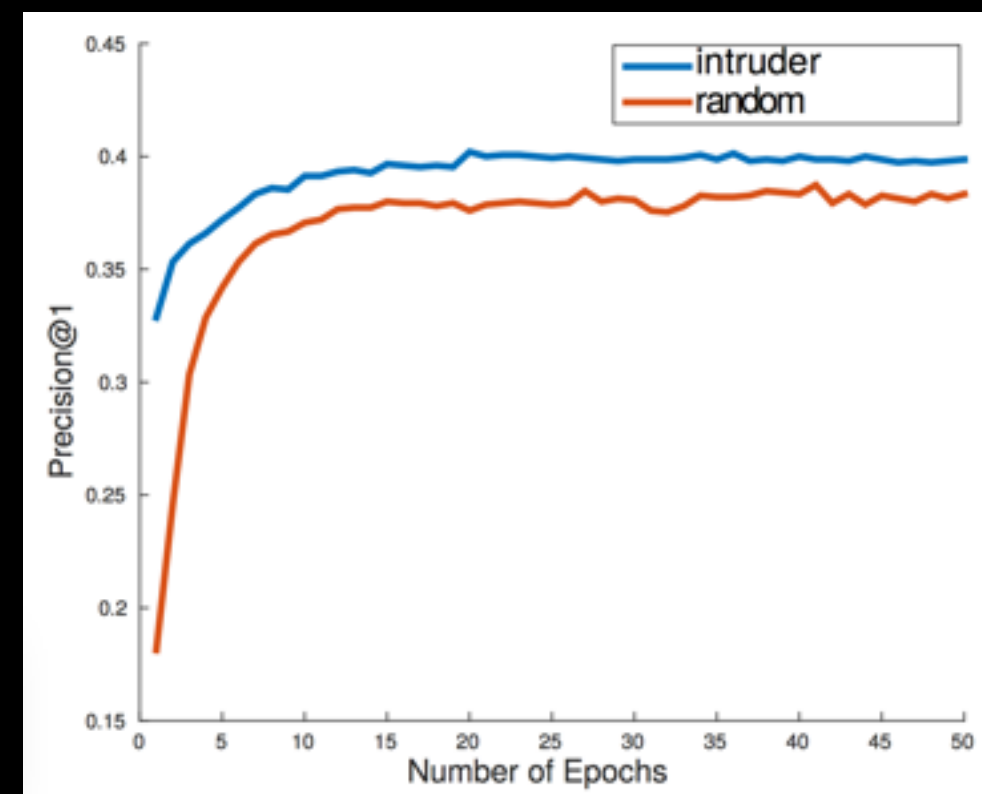
負例を賢く選択すると zero-shot learning の性能向上

表 11: 負例の選択方法 (ランダム vs ロス最小) の比較

	言語横断		クロスモーダル	
	ランダム	負例選択	ランダム	負例選択
P@1	38.4	40.2	3.7	5.6
P@5	54.2	55.5	10.9	12.4
P@10	60.4	61.8	15.8	17.8

(最大マージン法、画像キメラ)

(図3=右図) 少ない
epoch 数でも高い性能を
得ることができる



まとめと今後の課題

- ・ 線形の zero-shot learning における数学的・経験的な性質を考察し、性能を向上させる3つの手法（最大マージン化・データ拡張・負例選択）を提案した
- ・ 言語横断タスク、クロスモーダルタスクのいずれにおいても、state-of-the-art を大きく上回る性能を示した
- ・ 画像キメラと負例選択に、今後は意味を考慮した手法を適用したい