The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

カーネル法に基づく疎な言語表現のための高速計算可能な共起尺度

A Linear-Time Kernelized Co-occurrence Norm for Sparse Linguistic Expressions

横井 祥 *1*2 Sho Yokoi　小林 颯介 *3 Sosuke Kobayashi　福水 健次 *4 Kenji Fukumizu　乾 健太郎 *1*2 Kentaro Inui

*1東北大学 Tohoku University　*2理研 AIP センター RIKEN Center for AIP　*3Preferred Networks Preferred Networks　*4統計数理研究所 The Institute of Statistical Mathematics

Estimating pointwise mutual information (PMI), a well-known co-occurrence measure between linguistic expressions, leads to a trade-off between learning time and the robustness to data sparsity. We propose a new kernel-based co-occurrence measure, named pointwise HSIC (PHSIC). PHSIC, intuitively, is a "smoothed PMI" by kernels, so it is robust to data sparsity; furthermore, its estimator is reduced to an efficient linear-time matrix calculation. In our experiments, we apply PHSIC to a dialogue response selection task using sparse language data. Experimental results show that the learning speed is about 100 times faster than that of a recurrent neural network-based PMI estimator; moreover, when the size of the data is small, its predictive performance hardly deteriorates compared to PMI.

1. はじめに

ふたつの言語表現の間の共起の強さのモデル化は自然言語処理における基本的タスクである．たとえばコロケーション獲得 [Manning 99] では，はじめにコーパスから bigram（連続して出現する単語ペア）が集められ，とくに強く共起する bigram（たとえば "New York"）を見つけることが目的となる．あるいは対話の応答文選択 [Lowe 15] では，はじめに〈発話文，応答文〉ペアの集合が与えられ，あたらしい発話文 x に対して最も適当な応答文 y—発話文 x ともっとも強く共起する応答文 y—を候補集合から選択することが目的となる．いずれの場合も，はじめに言語表現のペアの集合 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ が与えられ，新しいペア (x, y) の共起の強さを計算・推定することになる．

言語表現のペアの共起の強さはしばしば**自己相互情報量（PMI）** [Church 89] でモデル化される．すなわち，与えられたペアの集合を適当な同時分布 $\mathbf{P}_{XY}$ からのサンプル

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \underset{\text{i.i.d.}}{\sim} \mathbf{P}_{XY} \tag{1}$$

だと考えて，ペア (x, y) の共起の強さを次で測る：

$$\mathrm{PMI}(x, y; X, Y) = \log \frac{\mathbf{P}_{XY}(x, y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)}. \tag{2}$$

PMI を与えられたデータ $\mathcal{D}$ から推定する方法をふたつ述べる．ひとつめは頻度による推定である：

$$\widehat{\mathrm{PMI}}_{\mathrm{MLE}}(x, y; \mathcal{D}) = \log \frac{n \cdot c(x, y)}{\sum_{y'} c(x, y') \sum_{x'} c(x', y)}. \tag{3}$$

ここで $c(x, y)$ はペア (x, y) の $\mathcal{D}$ 内の頻度を表す．式 (3) の推定は計算が容易であり，コロケーション獲得 *1 など，特に単語同士の共起を測る際にしばしば用いられるが [Church 89]，各 x や y が疎である場合（たとえば句や文である場合）は適用できない．ふたつめの PMI 推定法は再帰ニューラルネット（RNN）を用いるものである．Li ら [Li 16] は対話の応答文生成（各 x や y が文であり非常に疎な場合）に共起尺度として PMI を適用するため *2，$\mathbf{P}(y)$ を RNN 言語モデル [Mikolov 10]，$\mathbf{P}(y|x)$ を条件付き RNN 言語モデル [Sutskever 14] でモデル化し，PMI を推定した：

$$\widehat{\mathrm{PMI}}_{\mathrm{RNN}}(x, y; \mathcal{D}) = \log \frac{\widehat{\mathbf{P}}_{\mathrm{CondRNN}}(y|x)}{\widehat{\mathbf{P}}_{\mathrm{RNN}}(y)}. \tag{4}$$

式 (4) の推定は各 x や y が疎な場合も適用できるが，学習コストは大きい．以上のように，頻度による PMI の推定 (3) と RNN による PMI の推定 (4) には疎性への頑健性と計算コストというトレードオフがある．我々が知る限り，疎な言語表現に適用できかつ高速に計算可能な共起の尺度は存在しない．

本稿では，カーネル法に基づく新しい共起尺度 **Pointwise HSIC（PHSIC）** を提案する．PHSIC は直感的には "カーネルでスムージングした PMI" であり，疎性に頑健かつ高速な推定が可能という特徴を持つ．まず 2. 節で提案尺度の形式的な定義を与える．PMI が「相互情報量への貢献度」であることに対応させ，PHSIC は「HSIC（カーネル法に基づく依存性尺度）への貢献度」として定義する．3. 節で PHSIC が疎性に頑健であることの直感的な説明を与える．PMI は新しいペアとすでに観測されたペアを完全一致で照合するが，PHSIC はこの照合をカーネル関数で平滑化した量と見ることができる．4. 節で，PHSIC の行列計算に帰着させた効率的な推定法を示す．5. 節で対話の応答文選択の実験をおこなう．PHSIC は RNN による PMI の推定に比べて，学習は約 100 倍高速であり，かつデータ数が少ないとき（データがさらに疎な場合）に予測精度の劣化が小さいことを示す．

問題設定・記号

X, Y をそれぞれ $\mathcal{X}, \mathcal{Y}$ に値をとる確率変数とする．本稿で扱うタスクは，入力として言語表現のペアの集合 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \underset{\text{i.i.d.}}{\sim} \mathbf{P}_{XY}$ を受け取り，与えられたペア $(x, y) \in \mathcal{X} \times \mathcal{Y}$ の "共起の強さ" を測るものである．提案手法はカーネル法に基づくため，カーネル法に関連する記号をいくつか用意する (表 1)．

連絡先: yokoi@ecei.tohoku.ac.jp

*1　コロケーション獲得において単純に頻度 $c(x, y) \propto \widehat{\mathbf{P}}(x, y)$ の大きな bigram を探すと，"of the" など高頻度な機能語のペアが上位に上がってきてしまい，連語を探す目的に合わない [Manning 99].

*2　対話の応答文生成において単純に条件付き確率 $\widehat{\mathbf{P}}(y|x)$ の高い y を探すと，"I don't know." など，発話文 x の内容に関わらず用いることのできるありきたり回答が上位に上がってきてしまう [Li 16].

表 1: 本稿で用いる記号.

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	$\mathcal{X}$ 上の正定値カーネル（直感的には類似度関数）
$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$	$\mathcal{Y}$ 上の正定値カーネル
$\phi(x) := k(x, \cdot)$	k が定める特徴空間における $x \in \mathcal{X}$ の特徴ベクトル
$\psi(y) := \ell(y, \cdot)$	ℓ が定める特徴空間における $y \in \mathcal{Y}$ の特徴ベクトル
$\widetilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	確率分布 X で中心化したカーネル k（$\widetilde{\ell}$ も同様）
$\widetilde{k}(x, x') := k(x, x') - \mathbf{E}_{x'}[(x, x')] - \mathbf{E}_x[(x, x')] + \mathbf{E}_{x,x'}[(x, x')]$	
$= (\phi(x) - \mathbf{E}_x[\phi(x)])^\top (\phi(x') - \mathbf{E}_x[\phi(x)])$	

表 2: 提案尺度 Pointwise HSIC（PHSIC）の位置づけ（2. 節）.

確率変数間の依存性の尺度 X と Y の依存の度合い （$\mathbf{P}_{XY}$ と $\mathbf{P}_X\mathbf{P}_Y$ の異なりの度合い）	サンプル間の共起の尺度 ペア (x, y) の 「X と Y の依存の度合い」への貢献度
相互情報量（MI） $\mathrm{MI}(X, Y) = \mathrm{KL}[\mathbf{P}_{XY} \| \mathbf{P}_X\mathbf{P}_Y]$	自己相互情報量（PMI） $\mathrm{PMI}(x, y; X, Y)$
$= \mathbf{E}_{(x,y)}\left[\log \frac{\mathbf{P}_{XY}(x, y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)}\right]$	$= \log \frac{\mathbf{P}_{X,Y}(x, y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)}$
HSIC $\mathrm{HSIC}(X, Y; k, \ell) = \mathrm{MMD}^2[\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y]$	Pointwise HSIC（PHSIC） $\mathrm{PHSIC}(x, y; X, Y, k, \ell)$
$= \mathbf{E}_{(x,y)}\left[\mathbf{E}_{(x',y')}[\widetilde{k}(x, x')\widetilde{\ell}(y, y')]\right]$	$= \mathbf{E}_{(x',y')}[\widetilde{k}(x, x')\widetilde{\ell}(y, y')]$
$= \mathbf{E}_{(x,y)}\left[(\phi(x) - m_X)^\top C_{XY}(\psi(y) - m_Y)\right]$	$= (\phi(x) - m_X)^\top C_{XY}(\psi(y) - m_Y)$

2. Pointwise HSIC

この節でカーネル法に基づく新しい共起尺度である Pointwise HSIC（PHSIC）の形式的な定義を与える．はじめに PMI が「ペアの相互情報量への貢献度」と捉えられることを述べ，次にこれに対応させ，提案尺度 PHSIC を「ペアの HSIC（カーネル法に基づく依存性尺度）への貢献度」として定義する (表 2)．すなわち PHSIC は一言で言えば “カーネル化した PMI” である．

2.1 相互情報量と自己相互情報量

まず**確率変数の依存性**の測りかたについて述べる．確率変数 X と Y が独立であるとは，同時分布 $\mathbf{P}_{XY}$ と周辺分布の積 $\mathbf{P}_X\mathbf{P}_Y$ が等しいことであった．したがって，ふたつの確率変数 X と Y がどの程度依存しているかは，なんらかの観点で「$\mathbf{P}_{XY}$ と $\mathbf{P}_X\mathbf{P}_Y$ がどの程度異なるか」を測ればよい．

確率変数 X と Y の**相互情報量** [Cover 06] は *3 ：

$$\mathrm{MI}(X, Y) := \mathrm{KL}[\mathbf{P}_{XY} \| \mathbf{P}_X\mathbf{P}_Y]. \tag{5}$$

ここで $\mathrm{KL}[\cdot\|\cdot]$ は確率分布間の KL ダイバージェンスを表す．すなわち相互情報量は $\mathbf{P}_{XY}$ と $\mathbf{P}_X\mathbf{P}_Y$ の違いを KL ダイバージェンスの観点で測った確率変数の依存性尺度である．

ところで相互情報量は，同時分布 $\mathbf{P}_{XY}$ による期待値の形，つまりペア $(x, y) \in \mathcal{X} \times \mathcal{Y}$ 全体での総和の形でも表現することができる：

$$\mathrm{MI}(X, Y) = \mathbf{E}_{(x,y)}\left[\log \frac{\mathbf{P}_{XY}(x, y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)}\right]. \tag{6}$$

網掛け部がまさに**自己相互情報量（PMI）**

$$\mathrm{PMI}(x, y; X, Y) = \log \frac{\mathbf{P}_{XY}(x, y)}{\mathbf{P}_X(x)\mathbf{P}_Y(y)} \tag{7}$$

であることに注意すると，$\mathrm{PMI}(x, y)$ は「ペア (x, y) の，相互情報量 $\mathrm{MI}(X, Y)$ への貢献度」と見ることができる．

2.2 HSIC と PHSIC

カーネル法に基づく依存性尺度である HSIC を紹介し，つぎに HSIC への貢献度として提案尺度である PHSIC を定める．

Hilbert–Schmidt independence criterion（HSIC）はカーネル法に基づく確率変数間の依存性尺度である [Gretton 05]：

$$\mathrm{HSIC}(X, Y; k, \ell) = \mathrm{MMD}^2_{k,\ell}[\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y]. \tag{8}$$

ここで $\mathrm{MMD}[\cdot, \cdot]$ は maximum mean discrepancy（MMD）[Gretton 12] と呼ばれる量で，カーネルが定める特徴空間上での分布間の距離を定める．すなわち HSIC は $\mathbf{P}_{XY}$ と $\mathbf{P}_X\mathbf{P}_Y$ の違いを MMD の観点で測った確率変数の依存性尺度である．

ところで HSIC も相互情報量と同様に同時分布 $\mathbf{P}_{XY}$ による期待値の形で表現することができる：

$$\mathrm{HSIC}(X, Y; k, \ell) = \mathbf{E}_{(x,y)}\left[\mathbf{E}_{(x',y')}[\widetilde{k}(x, x')\widetilde{\ell}(y, y')]\right] \tag{9}$$

$$= \mathbf{E}_{(x,y)}\left[(\phi(x) - m_X)^\top C_{XY}(\psi(y) - m_Y)\right] \tag{10}$$

$$m_X := \mathbf{E}_x[\phi(x)], \quad m_Y := \mathbf{E}_y[\psi(y)] \tag{11}$$

$$C_{XY} := \mathbf{E}_{(x,y)}\left[(\phi(x) - m_X)(\psi(y) - m_Y)^\top\right]. \tag{12}$$

式 (9) は HSIC のデータ空間での表現であり，式 (10) は特徴空間での表現である *4．相互情報量と PMI の関係と同様に，**Pointwise HSIC（PHSIC）**を式 (9)(10) の網掛け部により定義する：

$$\mathrm{PHSIC}(x, y; X, Y, k, \ell) := \mathbf{E}_{(x',y')}[\widetilde{k}(x, x')\widetilde{\ell}(y, y')] \tag{13}$$

$$= (\phi(x) - m_X)^\top C_{XY}(\psi(y) - m_Y). \tag{14}$$

すなわち $\mathrm{PHSIC}(x, y)$ は「ペア (x, y) の，$\mathrm{HSIC}(X, Y)$ への貢献度」である．

3. 「平滑化した PMI」としての PHSIC

PHSIC がデータの疎性に頑健な理由として，PHSIC は「PMI をカーネルで平滑化した量」だと捉えられることを述べる．まず，PMI の推定量 (3) は次のように書き換えられる：

$$\widehat{\mathrm{PMI}}_{\mathrm{MLE}}(x, y) = \log \frac{n \cdot \sum_{i=1}^n \mathbb{I}[x = x_i \wedge y = y_i]}{\sum_{i=1}^n \mathbb{I}[x = x_i] \sum_{i=1}^n \mathbb{I}[y = y_i]}. \tag{15}$$

ここで $\mathbb{I}[\mathrm{cond}]$ は指示関数を表し，条件 cond が満たされるときは 1 を，そうでないときは 0 を返す．式 (15) によれば，PMI の推定量は「与えられた (x, y) を観測済みの $(x_i, y_i) \in \mathcal{D}$ たちと順に照合し，(x, y) と (x_i, y_i) が完全に一致すれば加点，x 側もしくは y 側のどちらか一方のみが一致する場合は減点」を繰り返した量である (表 3，上側)．

一方，PHSIC(13) の推定量は次の通り：

$$\widehat{\mathrm{PHSIC}}_{\mathrm{kernel}}(x, y) = \frac{1}{n}\sum_{i=1}^n \widehat{\widetilde{k}}(x, x_i)\widehat{\widetilde{\ell}}(y, y_i). \tag{16}$$

関数 $\widehat{\widetilde{k}}(\cdot, \cdot)$，$\widehat{\widetilde{\ell}}(\cdot, \cdot)$ はデータで中心化した類似度関数である *5．(16) によれば，PHSIC の推定量は「与えられた (x, y) を観測

*3 相互情報量は，一般に $I(X; Y)$ と表記されるが，本稿では後述する $\mathrm{HSIC}(X, Y)$ との表記の一貫性のため $\mathrm{MI}(X, Y)$ で表わす．

*4 一般にカーネル法では，特徴空間における内積 $\phi(x)^\top\phi(x')$ をデータ空間におけるカーネル $k(x, x')$ で計算できることを用いて，データ空間側で統計処理をおこなう．

*5 $\widehat{\widetilde{k}}(x, x') := k(x, x') - \frac{1}{n}\sum_{j=1}^n k(x, x_j) - \frac{1}{n}\sum_{i=1}^n k(x_i, x') + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i, x_j)$，$\widehat{\widetilde{\ell}}(\cdot, \cdot)$ も同様．

表 3: PMI と PHSIC の「与えられた (x, y) と観測済みの $(x_i, y_i) \in \mathcal{D}$ たちとの照合の仕方」という観点での比較.

	加点	減点
PMI	(x, y) ‖ ‖ $\mathcal{D} = \{\ldots, (x_i, y_i), \ldots\}$	(x, y) ‖ ≠ (x, y) ≠ ‖ $\{\ldots, (x_i, y_i), \ldots, (x_i, y_i), \ldots\}$
PHSIC	(x, y) ≈ ≈ (x, y) ≈ ≈ $\{\ldots, (x_i, y_i), \ldots, (x_i, y_i), \ldots\}$	(x, y) ≈ ≈ (x, y) ≈ ≈ $\{\ldots, (x_i, y_i), \ldots, (x_i, y_i), \ldots\}$

済みの $(x_i, y_i) \in \mathcal{D}$ たちと順に照合し，x 側と y 側がともに平均より似ている場合（$\widehat{\widetilde{k}}(x, x_i) > 0$ かつ $\widehat{\widetilde{\ell}}(y, y_i) > 0$ の場合）に加点 *6，x 側と y 側のどちらか一方が似ていてもう一方が似ていない場合は減点」を繰り返した量である (表 3，下側).

以上のように，PMI と PHSIC の推定量を「与えられた (x, y) と観測済みの $(x_i, y_i) \in \mathcal{D}$ たちとの照合の仕方」という観点で比較すると，PMI は完全一致で照合する一方，PHSIC はこれをカーネル（類似度関数）を用いて平滑化していることが分かる．PHSIC が（ナイーブな推定であってもなお）データの疎性に頑健であることがこの点からも期待できる．

3.1 既存の類似度関数の活用

また自然言語処理には正定値カーネルの形の類似度関数が潤沢に存在する．PHSIC はこれらの資産をそのまま活用できる．

- **単語ベクトル**の和 [Mikolov 13] や**文の分散表現** [Kiros 15] など，分布仮説に基づく文や句の分散表現が多数提案されており，学習済みのモデルも種々配布されている．さらに，これら同士の**コサイン**が簡便で高精度な類似尺度として利用できる．
- SVM 時代に考案された**構造カーネル** [Collins 02] が多数存在する．
- 正定値カーネルの和や積も正定値カーネルであるため，これらの類似度関数を自在に**組合せ**ることができる．

4. PHSIC の推定法と計算量

特徴空間での推定（線形カーネル・コサインの場合）：一般に，カーネル $k(\cdot, \cdot)$ が定める特徴ベクトル $\phi(\cdot)$ は未知ないし超高次元であり特徴空間での推定は難しい（これをデータ空間で代替するのがカーネル法の特徴のひとつであった）．ところが，昨今の自然言語処理で利用される句や文の類似度は，コサインでよく計算できるケースがある (3.1 節)．コサインとともに用いられる文 x のベクトル表現を $\boldsymbol{x} \in \mathbb{R}^d$ とすれば，これは特徴空間での表現 $\phi(x) = \boldsymbol{x}/\|\boldsymbol{x}\|_2$ が陽に与えられていることに等しい ($k(x, x') = \phi(x)^\top \phi(x') = \cos(\boldsymbol{x}, \boldsymbol{x}')$)．線形カーネルの場合も自明に同様の表現が手に入る．このような場合，PHSIC を特徴空間側で推定することが可能である．

PHSIC の特徴空間での表現 (14) の推定量は次の通り：

$$\widehat{\mathrm{PHSIC}}_{\mathrm{feat}}(x, y) = (\phi(x) - \overline{\phi(x)})^\top \widehat{C}_{XY} (\psi(y) - \overline{\psi(y)}) \quad (17)$$

$$\widehat{C}_{XY} := \tfrac{1}{n} \textstyle\sum_{i=1}^n \phi(x_i)\psi(y_i)^\top - \overline{\phi(x)}\,\overline{\psi(y)}^\top. \quad (18)$$

ただし $\overline{\phi(x)} = \frac{1}{n}\sum_{i=1}^n \phi(x_i)$ は特徴空間での標本平均，$\overline{\psi(y)}$ も同様．式 (17)(18) の計算は行列計算に帰着されており，しかも計算量も小さい．学習時はベクトル $\overline{\phi(x)}, \overline{\psi(y)} \in \mathbb{R}^d$ および行列 $\widehat{C}_{XY} \in \mathbb{R}^{d \times d}$ の計算が必要であり，時間計算量は $\mathcal{O}(nd^2)$，空間計算量は $\mathcal{O}(nd)$（ともにデータ数に対して線形）．予測時は $\phi(x), \psi(y) \in \mathbb{R}^d$ の構成および式 (17) の計算に $\mathcal{O}(d^2)$ の時間計算量を要する．

データ空間での推定（非線形カーネルの場合）：木カーネル [Collins 02] 等の非線形カーネルを用いる場合は PHSIC をデータ空間で推定する必要がある．データ空間でも，グラム行列の分解を介する簡単な工夫で式 (17) と同等の計算量での PHSIC の推定が可能である（紙面の都合で導出は省略する）．

$$\widehat{\mathrm{PHSIC}}_{\mathrm{ICD}}(x, y) = (\boldsymbol{a} - \overline{\boldsymbol{a}})^\top \widehat{C}_{\mathrm{ICD}} (\boldsymbol{b} - \overline{\boldsymbol{b}}) \quad (19)$$

$$\widehat{C}_{\mathrm{ICD}} := \tfrac{1}{n}(HA)^\top B. \quad (20)$$

ただし行列 $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)^\top \in \mathbb{R}^{n \times d}$ は不完全コレスキー分解により計算した $AA^\top \approx K$ *7 なる行列であり，B も同様 ($BB^\top \approx L$)．ベクトル $\overline{\boldsymbol{a}} \in \mathbb{R}^d$ は $\{\boldsymbol{a}_i\}_{i=1}^n$ の平均 $\frac{1}{n}A^\top \mathbf{1}$，$\overline{\boldsymbol{b}}$ も同様．また，新しい $x \in \mathcal{X}$ に対応するベクトル $\boldsymbol{a} \in \mathbb{R}^d$ は，不完全コレスキー分解のアルゴリズムを「途中から行う」ことで計算できる．分解中に採用された支配的な x_i たちを順に $x^{(1)}, \ldots, x^{(d)}$ とすると，$\boldsymbol{a}$ の第 j 成分 $\boldsymbol{a}(j)$ は以下で計算できる（$\boldsymbol{b}$ も同様）：$\boldsymbol{a}(j) = [k(x, x^{(j)}) - \sum_{m=1}^{j-1} \boldsymbol{a}(m) A_{jm}] \,/\, A_{jj}$.

5. 実験

PHSIC が疎なデータに適用でき (3. 節) しかも計算量が小さい (4. 節) ことを，疎な言語データ（文対）を用いる対話の応答文選択タスクを通して大規模実データで検証する．実験の結果，提案尺度は RNN を用いた PMI の推定に比べ 100 倍程度高速に学習でき，また訓練データ数が少ない場合（データがより疎な場合）も予測精度の劣化が少ないことがわかった．

タスク：対話の応答文選択タスクをおこなう．訓練データとして〈発話文，応答文〉のペアの集合 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ が与えられ，評価データの発話文 x 毎に応答文 y を選択肢から選ぶ．

データ：ツイッターにおける対話ペアを用いる．

- 検証・評価には，関連度（共起の度合い）が高い対話を人手で抽出したデータセットを用いる [Sordoni 15]*8.
- 訓練セットもやはり [Sordoni 15] と同様の方法で収集した．ただし，検証・評価セット（2012 年のツイート）とのデータの重複を避けるため，収集対象の年次を 2017 年に変更した．訓練セットの大きさ n の違いが学習時間・予測精度へ与える影響を確かめるため，収集したデータから乱択した $\{10^2, 10^3, 10^4, 10^5\}$ ペアを訓練セットとして実験をおこなう．
- 最後に，[Li 16] と同様に，〈文脈，発話，応答〉の三つ組を〈文脈＋発話，応答〉のペアに変換する．なお評価セットは各発話文 x に対してコーパス全体から乱択で選択された負例の応答 y' が 9 件ずつ割り当てられている（10 択問題となる）．

5.1 提案尺度：PHSIC

PHSIC は任意の正定値カーネルを用いることができる．今回は，もっとも簡単な類似尺度のひとつである，文を構成する単語ベクトルの和同士のコサイン（3.1 節）

$$k(x, x') = \cos(\textstyle\sum \mathrm{wordvecs}(x), \sum \mathrm{wordvecs}(x')) \quad (21)$$

を用いる．$\ell(\cdot, \cdot)$ も同様．単語ベクトルは，文字 n-gram を活用した単語ベクトル学習アルゴリズムである `fastText`

*6 x 側と y 側がともに平均より似ていない場合，すなわち $\widehat{\widetilde{k}}(x, x_i) < 0$ かつ $\widehat{\widetilde{\ell}}(y, y_i) < 0$ の場合も PHSIC の推定値が増加する．

*7 行列 $K \in \mathbb{R}^{n \times n}$ は $K_{ij} := k(x_i, x_j)$ なるグラム行列，$L_{ij} := \ell(y_i, y_j)$ も同様．

*8 `https://www.microsoft.com/en-us/download/details.aspx?id=52375`

モデル				訓練セットサイズ n			
				10^2	10^3	10^4	10^5
RNN	300	`fastText`	Total	11.7	20.6	99.2	634.3
			$\widehat{\mathbf{P}}(y)$	4.8	8.0	23.6	294.6
			$\widehat{\mathbf{P}}(y\|x)$	6.9	12.6	75.6	339.7
	1200	random	Total	31.3	65.5	161.5	1743.2
			$\widehat{\mathbf{P}}(y)$	8.6	21.7	57.4	675.6
			$\widehat{\mathbf{P}}(y\|x)$	22.7	43.8	104.1	1067.6
	1200	`fastText`	Total	33.8	49.0	162.0	1751.3
			$\widehat{\mathbf{P}}(y)$	10.8	16.3	57.2	671.0
			$\widehat{\mathbf{P}}(y\|x)$	23.0	32.7	104.8	1080.3
PHSIC	`fastText`		Total	**0.02**	**0.2**	**1.4**	**14.5**
			feat map	0.0	0.1	1.3	14.0
			mat calc	0.0	0.0	0.1	0.5

表 4: 各モデルの**学習時間**，単位は秒．各行はモデル，各列は訓練データサイズ n，太字は列毎の最短学習時間．RNN の各モデルに付記されているのは次元数および単語埋め込みの初期化法．PHSIC の計算時間は，内数として特徴関数の計算時間 (feat map) と行列計算時間 (mat calc) を付記した．

[Bojanowski 17] によりウェブクロールデータ *9 から学習済みの公開資源 *10 を利用する．ベクトルは 300 次元，語彙サイズは 200 万．

5.2 ベースライン尺度：RNN 言語モデル

訓練データで RNN 言語モデル [Mikolov 10] および条件付き RNN 言語モデル [Sutskever 14] を，1 層 LSTM[Hochreiter 97]，フレームワークは `Chainer`[Tokui 15] を用いて学習し，(1) 応答文の尤度 $\widehat{\mathbf{P}}(y)$，(2) 応答文の条件付き確率 $\widehat{\mathbf{P}}(y|x)$，および (3)PMI（$\log\left[\widehat{\mathbf{P}}(y|x)/\widehat{\mathbf{P}}(y)\right]$）をベースライン尺度として用いる．隠れ層の次元，および単語埋め込み層の初期化については，以下の設定の組合せで学習する：

- 隠れ層の次元：300，1200，
- 初期化：$[-0.1, 0.1]$ の一様ランダム，`fastText` (5.1 節).

ほか設定は以下の通り：

- 語彙：頻度 2 以上の語を利用，
- ドロップアウト率：0.1（300 次元），0.3（1200 次元），
- バッチサイズ：16（$n = 10^2$），64（$n = 10^3, 10^4, 10^5$），
- 最大エポック：30*11 ．

5.3 予測精度の評価

PHSIC やベースライン尺度は，各ペア (x, y) に対してなんらかの共起スコア $\mathrm{score}(x, y)$ を与える．共起スコアの良さを，(1)`MRR`（平均逆順位），(2)`AUC-ROC`，(1)`Recall@2` でそれぞれ評価する．

5.4 実験結果：学習時間

学習時間に関する実験結果は表 4 の通り *12．訓練セットの大きさ n が 10^5 のとき，PHSIC は RNN に基づく PMI の約 40–120 倍高速に学習が完了した．

5.5 実験結果：予測精度

予測精度に関する実験結果は表 5 の通り．10^5 万件の学習データをすべて使ったとき，PHSIC は既存尺度と大凡同等の予測精度を示し，さらに，訓練セットのサイズを小さくしたとき（より疎なとき），既存尺度の予測精度は大幅に劣化するのに対し，PHSIC の予測精度の劣化スピードははるかに緩慢であった．

モデル			訓練セットサイズ n			
			10^2	10^3	10^4	10^5
Chance Level			.50; .29; .20	.50; .29; .20	.50; .29; .20	.50; .29; .20
$\widehat{\mathbf{P}}(y)$	1200	`fastText`	.51; .30; .23	.50; .29; .21	.50; .30; .21	.50; .30; .21
$\widehat{\mathbf{P}}(y\|x)$	1200	`fastText`	.51; .30; .22	.50; .29; .21	.50; .30; .21	.52; .31; .23
PMI	300	random	.50; .29; .18	.51; .30; .21	.51; .30; .22	.58; .35; .29
		`fastText`	.50; .29; .18	.51; .29; .20	.56; .34; .25	**.66**; **.41**; **.36**
	1200	random	.50; .29; .20	.50; .29; .20	.51; .30; .19	.57; .35; .29
		`fastText`	.50; .29; .19	.51; .30; .20	.52; .32; .23	**.65**; **.42**; **.36**
PHSIC	`fastText`		**.57**; **.35**; **.29**	**.61**; **.38**; **.33**	**.62**; **.40**; **.34**	.62; **.40**; **.34**

表 5: 各モデルの**予測精度**（`MRR`，`ROC-AUC`，`Recall@2`）.

謝辞 東京大学の胡緯華氏からの有益な助言に感謝する．本研究の一部は文部科学省科研費 15H01702 の支援を受けた．

参考文献

[Bojanowski 17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.: Enriching Word Vectors with Subword Information, *TACL*, Vol. 5, pp. 135–146 (2017)

[Church 89] Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, in *ACL*, pp. 76–83 (1989)

[Collins 02] Collins, M. and Duffy, N.: Convolution Kernels for Natural Language, in *NIPS*, pp. 625–632 (2002)

[Cover 06] Cover, T. M. and Thomas, J. A.: *Elements of Information Theory*, John Wiley & Sons (2006)

[Gretton 05] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B.: Measuring Statistical Dependence with Hilbert-Schmidt Norms, in *ALT*, pp. 63–77 (2005)

[Gretton 12] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A.: A Kernel Two-Sample Test, *JMLR*, Vol. 13, No. Mar, pp. 723–773 (2012)

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (1997)

[Kiros 15] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S.: Skip-Thought Vectors, in *NIPS*, pp. 3294–3302 (2015)

[Li 16] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models, in *NAACL-HLT*, pp. 110–119 (2016)

[Lowe 15] Lowe, R., Pow, N., Serban, I., and Pineau, J.: The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems, in *SIGDIAL*, pp. 285–294 (2015)

[Manning 99] Manning, C. D. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, MIT press (1999)

[Mikolov 10] Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., and Khudanpur, S.: Recurrent neural network based language model, in *Interspeech*, pp. 1045–1048 (2010)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in *NIPS*, pp. 3111–3119 (2013)

[Sordoni 15] Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B.: A Neural Network Approach to Context-Sensitive Generation of Conversational Responses, in *NAACL-HLT*, pp. 196–205 (2015)

[Sutskever 14] Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, in *NIPS*, pp. 3104–3112 (2014)

[Tokui 15] Tokui, S., Oono, K., Hido, S., and Clayton, J.: Chainer: a Next-Generation Open Source Framework for Deep Learning, in *LearningSys* (2015)

*9 `http://commoncrawl.org/`

*10 `https://fasttext.cc/docs/en/english-vectors.html`

*11 最大エポック終了時まで，0.5 エポック（$n = 5 \times 10^5$）または 1 エポック（ほか）毎に検証セットパープレキシティを計測し，検証セットパープレキシティが最良のモデルでの予測精度，および（最大エポックまでに要した時間ではなく）当該モデルを獲得するまでの学習時間を報告する．

*12 計算機環境は，CPU：Xeon E5-1650-v3 (3.5GHz, 6 Cores)，GPU：GTX 1080 (8GB).