2B3-OS-07a-1in2

独立性尺度に基づく知識の粒度の教師なし推定

Unsupervised Granularity Estimation by Kernel Dependence Maximization

横井 祥 *1 Sho Yokoi　持橋 大地 *2 Daichi Mochihashi　高橋 諒 *1 Ryo Takahashi　岡崎 直観 *1 Naoaki Okazaki　乾 健太郎 *1 Kentaro Inui

*1東北大学 Tohoku University　*2統計数理研究所 The Institute of Statistical Mathematics

Modeling the association between items in a dataset is a problem that is frequently encountered in data and knowledge mining research. Most previous studies have simply applied a predefined fixed pattern to extract the substructure of each item pair and then analyzed the association between these substructures. The use of such fixed patterns may not, however, capture the significant association. To address this problem, we propose a novel machine learning task of extracting a strongly associated substructure pair (co-substructure) from each input item pair. We formalize it as a dependence maximization problem. Then, we discuss two critical issues in the task, namely the data sparsity problem and a huge search space. To address the data sparsity problem, we adopt the Hilbert–Schmidt independence criterion as an objective function. To improve search efficiency, we adopt the Metropolis–Hastings algorithm. We report results of empirical evaluations, in which the proposed method is applied to the acquisition of narrative event pairs, a knowledge mining task that is an active area of study in the field of natural language processing.

1. はじめに

関係のあるオブジェクト対の獲得や，オブジェクト間の関係のモデル化は，データマイニング・機械学習の主要タスクのひとつである．たとえば自然言語処理においても，事態間関係（例：⟨*X commit a crime*, *X be arrested*⟩）の獲得・予測 [Chambers 08, etc.]，選択選好（述語動詞と項の相性の良さのモデル化）[Resnik 97, etc.]，など，多くの問題が関係のある言語表現対の収集や関係のモデル化を伴う．

関係知識の獲得と予測の典型的な 3 ステップを，Chambers らによる事態関係知識獲得を例に挙げて述べる [Chambers 08]．**Step 1.** はじめに，関係を持つオブジェクト対を収集する．例えば事態関係知識獲得においては，コーパスから共参照項を持つ文対（例：⟨*Tom*$_i$ *killed Nancy.*, *The police arrested Tom*$_i$ *immediately.*⟩）を収集する．**Step 2.** 次にオブジェクト対を抽象表現に変換する．例えば述語動詞および注目している登場人物の係り受けパス（例：subject, object）に着目して，収集した文対を抽象表現（例：⟨*X kill*, *arrest X*⟩）に変換する．**Step 3.** 最後にオブジェクトのペアに対するスコアリング方法を与える（これは新しいペアに対して関係の有無を判定する方法でもある）．たとえば Chambers らは，自己相互情報量 (PMI) により事態間関係をスコアリングする．

Step 2 では，従来，事前に定義されたテンプレートないし特徴量を用いてオブジェクト対を抽象化してきた．しかし固定された抽象表現ではオブジェクト対が持つ関係を適切に捉えられない場合がある．例えば事態間関係知識において，抽象表現に含める語により関係知識の意味が大きく変化する [Granroth-Wilding 16]．たとえば前述の手法では，文対 ⟨*X had had absent repeatedly.*, *X was fired.*⟩ は ⟨*X have*, fire *X*⟩ に抽象化され，文対 ⟨*X has a talent for accounting work.*, *X was hired with favorable treatment.*⟩ は ⟨*X have*, hire *X*⟩ に抽象化されるが，⟨*X have*, fire *X*⟩ と ⟨*X have*, hire *X*⟩ は明らかに矛盾している．この場合，⟨*X have absent repeatedly*, fire *X*⟩ と ⟨*X have talent*, hire *X*⟩ が獲得されることが望ましい．

連絡先: yokoi@ecei.tohoku.ac.jp

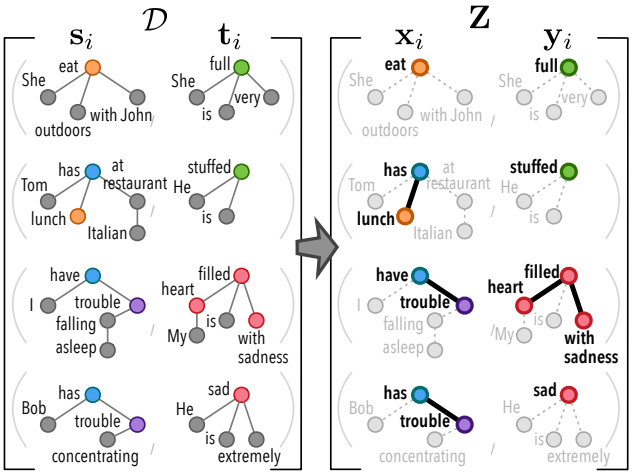


図 1: 部分構造対抽出問題の入出力．

本稿でははじめに，オブジェクト対 ⟨$\mathbf{s}_i$, $\mathbf{t}_i$⟩ 毎に，その関係に寄与る部分構造対 ⟨$\mathbf{x}_i$, $\mathbf{y}_i$⟩ を従属性最大化の観点で抽出する新しい機械学習のタスク（図 1）を提案する．次に，目的関数（従属性尺度）として Hilbert–Schmidt independence criterion を，探索方法としてメトロポリス・ヘイスティングス法に基づくサンプリングを採用した提案手法を示す．最後に，小規模人工データを用いた実験により提案手法が望ましい挙動を示すことを確認し，実コーパスを用いた事態間関係の予測タスクに提案手法を適用することでインスタンス毎の抽象表現抽出が予測精度という観点でも効果的であることを示す．

2. 部分構造対抽出

はじめにインスタンス・ペア毎に部分構造を抽出する問題を定式化する．

部分構造対抽出．オブジェクト対それぞれが持つ関連の強い部分構造対（ペアをペアたらしめる部分構造）を従属性最大化の観点で抽出する．

入力： オブジェクト対の集合 $\mathcal{D} = \{(\mathbf{s}_i, \mathbf{t}_i)\}_{i=1}^N$．$\mathbf{s}_i \in \mathcal{S}$, $\mathbf{t}_i \in \mathcal{T}$．各ペア $(\mathbf{s}_i, \mathbf{t}_i)$ は何らかの関係（共参照関係，共起関係など）を持つ．

出力： 従属性（後述）を最大化するような，オブジェクトの部分構造対の集合 $\mathbf{Z} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$．ここで $\mathbf{x}_i \preccurlyeq \mathbf{s}_i$, $\mathbf{y}_i \preccurlyeq \mathbf{t}_i$ であり，'$\preccurlyeq$' は何らかの部分構造（部分集合，部分木など）を表す．従属性は，$\mathbf{Z}$ を適当な同時分布からの独立な N サンプル

$$\mathbf{Z} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \sim P_{XY}, \tag{1}$$

と見なしたときの確率変数 X と Y の従属性，すなわち同時分布 P_{XY} と周辺分布の積 $P_X P_Y$ の何らかの尺度による距離によって推定する．

図 1 は部分構造対抽出の問題の理想的な入出力である．句同士（たとえば *heart filled with sadness* と *sad*）の類似性が考慮されながら，イベントペアをペアたらしめる（たとえば因果関係を特徴付ける）部分構造が抽出されている．

従属性の尺度，すなわち P_{XY} と $P_X P_Y$ の距離尺度として，データマイニング・機械学習の文脈では典型的には相互情報量 (MI) が用いられる [Church 90, Peng 05]．相互情報量を用いると $\mathbf{Z} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ の従属性は

$$\mathrm{MI}(\mathbf{Z}) = \sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) p(\mathbf{y})} \tag{2}$$

$$= \mathrm{KL}[P_{XY} \| P_X P_Y]. \tag{3}$$

$\mathrm{KL}[\cdot\|\cdot]$ はカルバック・ライブラー・ダイバージェンス．

ところが，頻度を用いたナイーブな相互情報量の推定を用いて部分構造抽出の問題を解こうとすると，ふたつの大きな問題が生じる：**データ・スパースネス**．解 $\mathbf{Z}$ には，$\mathbf{x}_i$, $\mathbf{y}_i$ として $\mathbf{s}_i$, $\mathbf{t}_i$ 任意の大きさの部分構造（たとえば "*She has big dinner*"）が含まれるため，頻度ベースの推定はゼロ頻度問題に直面する．**巨大な探索空間**．部分構造対抽出の問題の解空間は，部分構造 $\mathbf{x}_i$, $\mathbf{y}_i$ の取り方の組合せの全体であり，全探索は極めて困難である（組合せ爆発）．

3. 提案手法

データ・スパースネス問題と巨大な探索空間に対処するため，部分構造対抽出の問題の目的関数としてカーネル法ベースの従属性尺度 HSIC を，探索に MCMC サンプリングを用いた確率的山登りを採用する．セクションの最後で提案手法の計算コストについて議論する．

3.1 目的関数：HSIC

スパースな部分構造の集合に対して従属性を推定するため，本稿ではカーネル法ベースの従属性（依存性）尺度である Hilbert–Schmidt independence criterion (HSIC) [Gretton 05] を採用する．HSIC は，特徴選択 [Song 12]，次元削減 [Fukumizu 09] など多くの機械学習・データマイニングの問題に適用されている，独立性・従属性の尺度である．

はじめに HSIC の推定量の計算方法を述べる．X, Y をそれぞれ $\mathcal{X}, \mathcal{Y}$ に値をとる確率変数，$\mathbf{Z} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \subseteq \mathcal{X} \times \mathcal{Y}$ を同時分布 P_{XY} からの独立な N サンプルとするとき，HSIC の推定量，すなわち X と Y の従属性の度合いは次のように計算できる：

$$\mathrm{HSIC}(\mathbf{Z}; k, \ell) = \frac{1}{N^2} \mathrm{tr}(\mathbf{KHLH}) = \frac{1}{N^2} \mathrm{tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}). \tag{4}$$

- $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ は正定値カーネル．直感的には，これらは部分構造間の類似度関数を表す．
- $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j)) \in \mathbb{R}^{N \times N}$, $\mathbf{L} = (\ell(\mathbf{y}_i, \mathbf{y}_j)) \in \mathbb{R}^{N \times N}$ はグラム行列．これらはカーネル関数 k, ℓ によるデータ $\{\mathbf{x}_i\}$, $\{\mathbf{y}_i\}$ それぞれの類似度行列を表す．
- $\tilde{\mathbf{K}} = \mathbf{HKH} \in \mathbb{R}^{N \times N}$, $\tilde{\mathbf{L}} = \mathbf{HLH} \in \mathbb{R}^{N \times N}$ は中心化グラム行列である．$\mathbf{H} = ((\delta_{ij} - \frac{1}{N})) \in \mathbb{R}^{N \times N}$．

直感的には，HSIC は部分構造間の類似度によってスムージングした相互情報量である．つまり，従属性を推定をする際に，たとえば $\mathbf{x}_i =$ "*have dinner*" と $\mathbf{x}_j =$ "*have lunch*" の類似度を考慮に入れることができる．このことを示すため，はじめに，データ $\{\mathbf{x}_n\} := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ よって "中心化したカーネル関数" $\tilde{k}(\cdot, \cdot; \{\mathbf{x}_n\})\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$，および "**カーネル *PMI (kPMI)***" $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ を次の通り定義する：

$$\begin{aligned} \tilde{k}(\mathbf{x}, \mathbf{x}'; \{\mathbf{x}_n\}) &:= k(\mathbf{x}, \mathbf{x}') - \frac{1}{N} \sum_{j=1}^N k(\mathbf{x}, \mathbf{x}_j) \\ &\quad - \frac{1}{N} \sum_{i=1}^N k(\mathbf{x}_i, \mathbf{x}') + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j), \end{aligned} \tag{5}$$

$$\mathrm{kPMI}(\mathbf{x}, \mathbf{y}; \mathbf{Z}) := \sum_{i=1}^N \tilde{k}(\mathbf{x}, \mathbf{x}_i; \{\mathbf{x}_n\}) \tilde{\ell}(\mathbf{y}, \mathbf{y}_i; \{\mathbf{y}_n\}). \tag{6}$$

$\tilde{\ell}(\cdot, \cdot; \{\mathbf{y}_n\})$ も $\tilde{k}$ と同様に定義する．これらを用いると，頻度を用いてナイーブに推定する PMI/相互情報量と kPMI/HSIC の間には，表 1 に示すような密な関係があることが分かる．PMI および相互情報量は他のデータを完全一致で参照するのに対し，kPMI および HSIC はカーネル関数 k, ℓ によって計算される他のデータとの類似性が考慮される．また，相互情報量は PMI の足し合わせであり，HSIC は kPMI の足し合わせである．すなわち HSIC は，周りのデータそれぞれとの類似性を考慮して "スムージングされた" 相互情報量と見ることができる．

HSIC の計算に用いる k, ℓ は正定値カーネルであれば何でも構わない．SVM などカーネル法ベースの手法とともに開発されてきた各種カーネルを利用することができる．

3.2 探索：MH

巨大な探索空間に対処するため，本稿ではマルコフ連鎖モンテカルロ法の一種であるメトリポリス・ヘイスティングス法 (MH) [Chib 95] を用いた確率的な山登りにより，最適解（に近い解）を探索する．はじめに次の確率分布を考える：

$$p(\mathbf{Z}; k, \ell, \beta) \propto \exp(\beta \cdot \mathrm{HSIC}(\mathbf{Z}; k, \ell)). \tag{7}$$

β は逆温度．この分布の上では，大きな HSIC 値を持つ（すなわち従属性が大きな）$\mathbf{Z}$ は大きな確率値を持ち，小さな HSIC 値を持つ（すなわち従属性が小さな）$\mathbf{Z}$ は小さな確率値を持つ．提案手法では，$\mathbf{Z}$ を少しずつ変化させつつ分布 (7) から $\mathbf{Z}$ のサンプリングを繰り返すことで，全探索よりも小さなコストで最適に近い解を探索する（図 2）．具体的な手順は以下の通り．

1. 現在の状態（解）を $\mathbf{Z} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ とする．

表 1: PMI/MI と kPMI/HSIC の関係．$\mathbb{I}$[condition] は条件が真のとき 1，さもなくば 0．‘$\veebar$’ は排他的論理和．中心化グラム行列 $\tilde{\mathbf{K}}$ の元は $\tilde{\mathbf{K}}_{ij} = \tilde{k}(\mathbf{x}_i, \mathbf{x}_j; \{\mathbf{x}_n\})$ で計算でき，HSIC は $\mathrm{HSIC}(\mathbf{Z}) = \frac{1}{N^2}\sum_{ij}\tilde{\mathbf{K}}_{ij}\tilde{\mathbf{L}}_{ij}$ であることを用いた．

	$(\mathbf{x},\mathbf{y})$ の $(\mathbf{x}_i,\mathbf{y}_i)\in\mathbf{Z}$ との整合性		$(\mathbf{x},\mathbf{y})$ の $\mathbf{Z}$ 全体との整合性	従属性の推定量
MI	+	$\mathbf{x}=\mathbf{x}_i \wedge \mathbf{y}=\mathbf{y}_i$	$\mathrm{PMI}(\mathbf{x},\mathbf{y};\mathbf{Z}) = \log \frac{N\cdot\sum_i \mathbb{I}[\mathbf{x}=\mathbf{x}_i \wedge \mathbf{y}=\mathbf{y}_i]}{\sum_i \mathbb{I}[\mathbf{x}=\mathbf{x}_i]\sum_i \mathbb{I}[\mathbf{y}=\mathbf{y}_i]}$	$\mathrm{MI}(\mathbf{Z}) = \frac{1}{N}\sum_i \mathrm{PMI}(\mathbf{x}_i,\mathbf{y}_i)$
	−	$\mathbf{x}=\mathbf{x}_i \veebar \mathbf{y}=\mathbf{y}_i$		
HSIC	+	$\tilde{k}(\mathbf{x},\mathbf{x}_i)\tilde{\ell}(\mathbf{y},\mathbf{y}_i) > 0$	$\mathrm{kPMI}(\mathbf{x},\mathbf{y};\mathbf{Z}) = \sum_i \tilde{k}(\mathbf{x},\mathbf{x}_i;\{\mathbf{x}_n\})\tilde{\ell}(\mathbf{y},\mathbf{y}_i;\{\mathbf{y}_n\})$	$\mathrm{HSIC}(\mathbf{Z}) = \frac{1}{N^2}\sum_i \mathrm{kPMI}(\mathbf{x}_i,\mathbf{y}_i)$
	−	$\tilde{k}(\mathbf{x},\mathbf{x}_i)\tilde{\ell}(\mathbf{y},\mathbf{y}_i) < 0$		

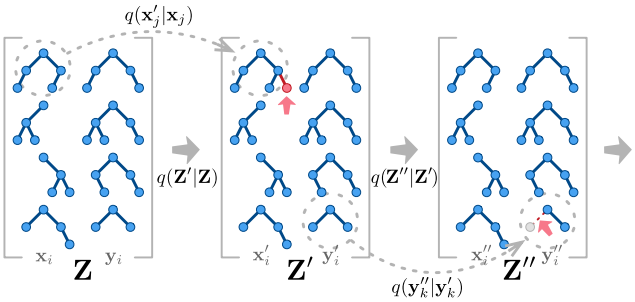


図 2: MH によるサンプリング．

2. $\mathbf{Z}$ の部分構造をひとつだけ変化させて次の状態の候補 $\mathbf{Z}'$ を作る．はじめに $\mathbf{x}_i$ または $\mathbf{y}_i$ を $\mathbf{Z} = \{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$ から一様分布で選択する：$\forall i,\ p(\mathbf{x}_i|\mathbf{Z}) = p(\mathbf{y}_i|\mathbf{Z}) = \frac{1}{2N}$．次に提案分布 $q(\mathbf{x}'_i|\mathbf{x}_i)$（実験で用いた具体的な分布は実験のセクションで述べる）により，$\mathbf{x}_i$ から $\mathbf{x}'_i$ を作る．したがって，$\mathbf{Z} = \{\ldots,(\mathbf{x}_i,\mathbf{y}_i),\ldots\}$ から $\mathbf{x}_i$ のみを変化させた $\mathbf{Z}' = \{\ldots,(\mathbf{x}'_i,\mathbf{y}_i),\ldots\}$ が次の候補となる確率は

$$q(\mathbf{Z}'|\mathbf{Z}) = q(\mathbf{x}'_i|\mathbf{x}_i)p(\mathbf{x}_i|\mathbf{Z}) = \frac{1}{2N}q(\mathbf{x}'_i|\mathbf{x}_i). \tag{8}$$

3. 候補 $\mathbf{Z}'$ を確率 $\min(1,r)$ で採択する．ここで r は

$$r = \frac{p(\mathbf{Z}';k,\ell,\beta)}{p(\mathbf{Z};k,\ell,\beta)}\cdot\frac{q(\mathbf{Z}|\mathbf{Z}')}{q(\mathbf{Z}'|\mathbf{Z})} \tag{9}$$

$$= \exp(\beta(\mathrm{HSIC}(\mathbf{Z}';k,\ell) - \mathrm{HSIC}(\mathbf{Z};k,\ell)))\frac{q(\mathbf{x}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x})}. \tag{10}$$

4. ステップ 2–3 を繰り返す．

3.3 計算コスト

一見すると，グラム行列の構成に $O(n^2)$，HSIC の推定に $O(n^3)$ のコストがかかるように見える．ところが実際には，グラム行列を構成しなければいけないのは MH の繰り返しの最初の 1 回目だけであり，その後は $\mathbf{Z}$ から $\mathbf{Z}'$ をサンプルする際に変化するただひとつの $\mathbf{x}_i$（または $\mathbf{y}_i$）に対応するグラム行列の行・列を $O(n)$ で更新するだけで良い．また，HSIC の推定も，グラム行列の不完全コレスキー分解，および分解された行列による HSIC の推定 [Gretton 05, Appendix 4] を介することで，$O(n\kappa^2)$ でおこなうことができる．κ は分解後の次元数．したがって，MH の繰り返し毎の HSIC の推定コストは $O(n\kappa^2)$ に抑えることができる．

4. 実験

イベントのペア [Chambers 08] の抽象表現の獲得，および，イベントのペアの予測を通して，部分構造対抽出の問題設定の妥当性と，提案手法の効果を検証する．

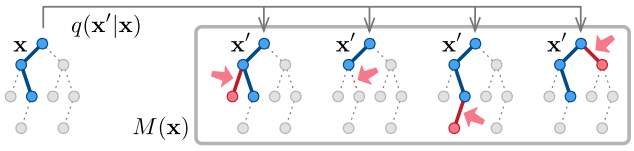


図 3: 提案分布 $q(\mathbf{x}'|\mathbf{x})$．

4.1 実験設定

4.1.1 データ構造

文 $\mathbf{s}_i$，$\mathbf{t}_i$ はそれぞれ依存構造木で，その部分構造 $\mathbf{x}_i$，$\mathbf{y}_i$ はその根付き部分木とする（図 1）．

4.1.2 カーネル関数

単語ベクトルから加法性に基づき部分構造のベクトルを計算し，そのコサイン [*1] を部分構造間の類似度とした：

$$k(\mathbf{x}_i,\mathbf{x}_j) = \cos(\mathrm{ave}(\mathrm{vecs}(\mathbf{x}_i)),\mathrm{ave}(\mathrm{vecs}(\mathbf{x}_j))). \tag{11}$$

$\ell(\cdot,\cdot)$ も同様．単語ベクトルは，Mikolov らの SGNS [Mikolov 13] で GoogleNews から学習済みの 300 次元のベクトルを用いた [*2]．

4.1.3 提案分布

提案分布は $q(\mathbf{x}'|\mathbf{x}) = 1/|M(\mathbf{x})|$ とした．$M(\mathbf{x})$ は，$\mathbf{x}\preccurlyeq\mathbf{s}$ に対して，$\mathbf{x}$ からただひとつだけ枝を伸ばしたり縮めたりして作ることのできる $\mathbf{s}$ の部分木の集合（図 3）．

4.2 実験 1：知識獲得

12 の人工的なデータに対して提案手法を適用した結果は図 4 の通り．以下の結果が見て取れる：

- ブロック毎（1–4, 2–8, 9–12）に共通項が残される．
- 1 回だけ出現する語（dinner, lunch）が breakfast との類似性により残される．
- 第 1 ブロックと第 2 ブロックの弁別を可能にすべく，(with) friend が残される．

類似性に基づいて意味によるスムージングが行われ，また相互情報量と同様 $\mathbf{x}\mapsto\mathbf{y}$，$\mathbf{y}\mapsto\mathbf{x}$ 両方向の予測が可能になる部分構造が出力されることが分かる．

4.3 実験 2：予測

先行研究 [Chambers 08, etc.] にのっとり，イベントの対が典型的な組合せであるかどうかのペア分類問題を解く．

4.3.1 コーパス

- **The Gigaword Corpus**[*3]，NYT，2000 年の記事．先行研究 [Chambers 08, Chambers 09, Granroth-Wilding 16] にのっとる．
- **Andrew Lang Fairy Tale Corpus**[*4] 全体．先行研

*1 $\cos(\cdot,\cdot)$ は正定値であり HSIC の適用条件を満たす．
*2 https://code.google.com/archive/p/word2vec/
*3 https://catalog.ldc.upenn.edu/ldc2003t05/
*4 http://www.mythfolklore.net/andrewlang/

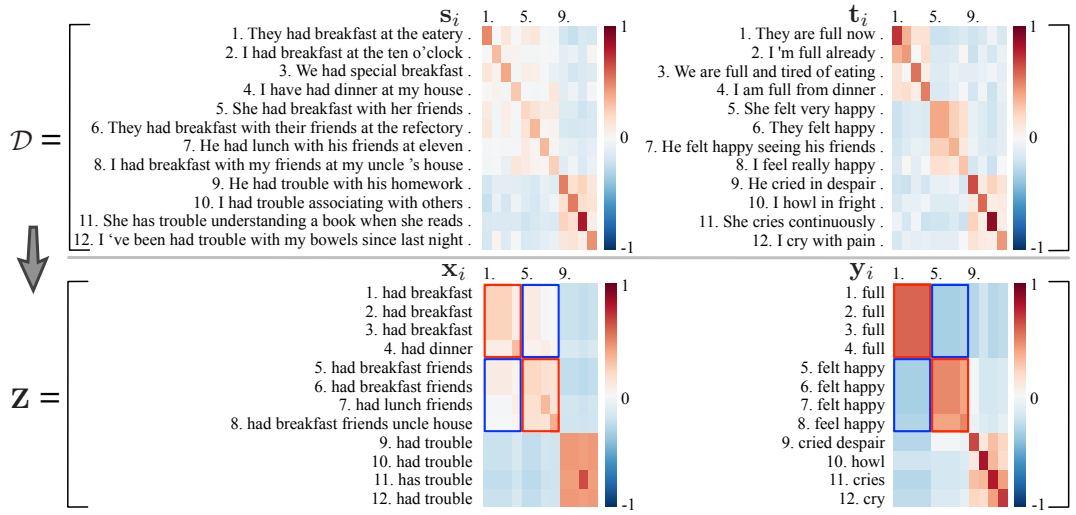
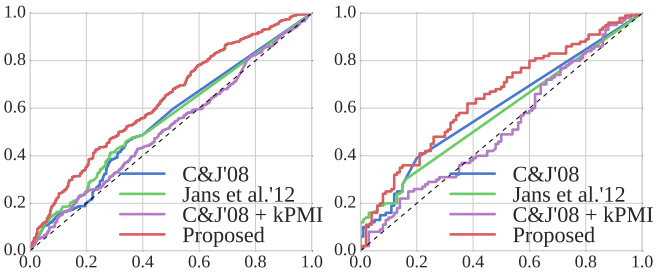


図 4: 予備実験の入力 $\mathcal{D} = \{(\mathbf{s}_i, \mathbf{t}_i)\}$，出力 $\mathbf{Z} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$，およびそれぞれに対応する中心化グラム行列のヒートマップ．



(a) Gigaword　　(b) Fairy Tale

図 5: 予測実験の結果：ROC カーブ．

表 2: 予測実験の結果：ROC-AUC 値．

Method	Abstraction	Model	Gigaword	Fairy Tale
C&J'08	Fixed (C&J)	PMI	0.553	0.596
Jans *et al.*'12	Fixed (C&J)	Conditional	0.556	0.576
C&J'08 + kPMI	Fixed (C&J)	kPMI	0.518	0.518
Proposed	**Dynamic**	kPMI	**0.633**	**0.646**

究 [Jans 12] にのっとる．

4.4 評価尺度

イベント・ペアの正例のスコアが負例のスコアより高くなるかどうかに興味があるので，ROC-AUC を採用する．

4.5 ベースライン手法

ベースラインとして，固定された抽象表現を用いてイベントペアのモデル化をする二手法を採用する [Chambers 08, Jans 12]．

4.6 実験結果

実験結果を図 5 および表 2 に示す．インスタンス毎の抽象表現の決定が予測精度に大きく貢献することが分かる．

参考文献

[Chambers 08] Chambers, N. and Jurafsky, D.: Unsupervised Learning of Narrative Event Chains, in *ACL*, pp. 789–797 (2008)

[Chambers 09] Chambers, N. and Jurafsky, D.: Unsupervised Learning of Narrative Schemas and their Participants, in *ACL*, pp. 602–610 (2009)

[Chib 95] Chib, S. and Greenberg, E.: Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, Vol. 49, No. 4, pp. 327–335 (1995)

[Church 90] Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational linguistics*, Vol. 16, No. 1, pp. 22–29 (1990)

[Fukumizu 09] Fukumizu, K., Bach, F. R., and Jordan, M. I.: Kernel dimension reduction in regression, *Annals of Statistics*, Vol. 37, No. 4, pp. 1871–1905 (2009)

[Granroth-Wilding 16] Granroth-Wilding, M. and Clark, S.: What Happens Next? Event Prediction Using a Compositional Neural Network Model, in *AAAI*, pp. 2727–2733 (2016)

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

[Jans 12] Jans, B., Bethard, S., Vulić, I., and Moens, M. F.: Skip N-grams and Ranking Functions for Predicting Script Events, *EACL*, pp. 336–344 (2012)

[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)

[Peng 05] Peng, H., Long, F., and Ding, C.: Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226–1238 (2005)

[Resnik 97] Resnik, P. S.: Selectional Preference and Sense Disambiguation, in *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pp. 52–57 (1997)

[Song 12] Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K.: Feature Selection via Dependence Maximization, *JMLR*, Vol. 13, pp. 1393–1434 (2012)