テキスト情報による関係の主-双対表現と共通空間射影法

大岩 秀和*
東京大学 情報理工学系研究科
hidekazu.oiwa@gmail.com

辻井 潤一
Microsoft Research
jtsujii@microsoft.com

1　はじめに

関係の連続値ベクトル表現に関する研究が近年盛んである．関係間の意味的類似度やフレーズの極性の判定方法として関係の連続値ベクトル表現は利便性等の面で有効であり，タスクに応じて様々な構成方法が提案されてきた [10, 16]．本稿では，関係マイニング・関係表現マイニングに有効な二項関係の連続値ベクトル表現の構成方法を提案し，その有効性を検証する．

本手法は二項関係が持つ双対性 [2] を利用する．ある関係性を内在するエンティティペアの部分集合によって，関係は外延的に構成出来る．一方で，関係はエンティティペアを結ぶ関係表現の部分集合で表現される．これらエンティティのペアとペア間の関係表現は，特定の関係に対応する部分集合がクラスタを形成するように，テキスト情報を用いてそれぞれ意味空間の構築が可能である．本稿では関係の二面性に着目し，この二つの意味空間を一つの共通空間へ射影する方法を提案する．共通空間への射影により，各意味空間で同じ関係を示すクラスタ同士を結びつけ互いに有用な情報を交換し，より豊富な情報を持つ共通空間を構成出来る．本提案手法の概要を図 1 で図示する．本稿では，共通空間の構成法と実験による本提案手法の関係マイニング・関係表現マイニングへの有効性を示す．

1.1　関係マイニングと関係表現マイニング

ある関係を持つエンティティペアを列挙するタスクを関係マイニング，関係表現を列挙するタスクを関係表現マイニングと呼ぶ．エンティティペア集合や関係表現集合を人手で全列挙することは困難なため，巨大コーパスからこれら集合を自動的に作成するマイニング方法が提案されてきた．関係表現マイニングは既存研究では関係抽出の補助タスクとして扱われることが多かった．共通意味空間の構築は，関係マイニングおよび関係表現マイニングに対して様々な利点をもつ．

これらのタスクの適合率と再現率は初めに与えられた集合の質と量に強く依存する．初期集合が小さい場合，再現率は低くなる．初期集合を大きくすると，曖昧な関係表現や不正確なエンティティペアが初期集合に混入する影響で適合率が低くなりやすい．従って，大量の曖昧性の低いオブジェクトの収集が重要になる．本稿で提案する共通空間射影法は，類似度計算の精度を高めるだけでなく，各オブジェクトの曖昧性を測るための重要な情報を提供する (2.5 節)．さらに，共通意味空間ではエンティティペアと関係表現は同じ連続値ベクトル表現を持つため，これらの間の類似度計算も自然と可能になる．そのため，少数のエンティティペアから意味的類似度の高い関係表現を集めるなど，マイニングを交互に行う事が可能になる (3.3 節)．

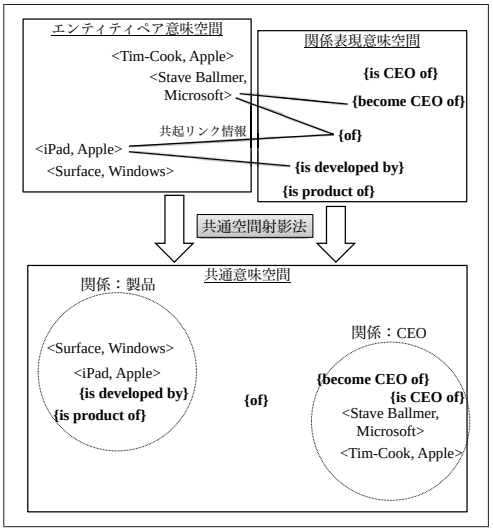

図 1: 共通意味空間への射影

1.2　関連研究

エンティティペアと関係表現の双対性に着目した既存研究として，[2] が挙げられる．この論文では，共起情報を用いてエンティティペアと関係表現を同時にクラスタリングする手法が提案された．我々の提案手法は，共起情報に加えエンティティペアと関係表現の特徴付けを利用して，より豊富な情報を持つ連続値ベクトル表現を構成する点において，[2] の拡張法とみなすことが出来る．また，共起エンティティペア情報を関係表現の特徴量として関係表現マイニングに用いた研究も存在する [9]．提案手法は，エンティティペアの特徴付けを通して関係表現間の特徴ベクトルを改良しているため，この既存研究の拡張ともみなせる．その他，関係マイニングや関係表現マイニングタスクのための様々なアルゴリズム [4, 5, 13, 15] が提案されているが，これらは我々の手法と異なり関係表現の定義に強い制限が存在する．さらに，関係の双対性には着目していないため，エンティティペアと関係表現間の類似度を測る事もできない．

*日本学術振興会特別研究員．本稿は，第一著者が Microsoft Research Asia 訪問時の研究成果である．

2 基本フレームワーク

2.1 外延集合と表明集合

E をエンティティの集合と定義する．エンティティのペアを $\langle e_i, e_j \rangle$ と表記し $(e_i, e_j \in E)$，エンティティペアの集合を E^2 と定義する．関係は R で表記し，エンティティペアの部分集合 $E_R \subset E^2$ によって外延的に定義される．例として，関係 CEO は $\{ \langle$Tim-Cook, Apple$\rangle, \langle$Ballmer, Microsoft$\rangle, \ldots\}$ で外延的に定義出来る．本稿ではこの集合を関係 R の外延集合と呼ぶ．

一方で，関係 R はテキスト中の様々な関係表現によって表明される．関係表現は r_i と表記し，関係表現の集合を D で定義する．ある関係 R を表明する関係表現の部分集合を関係 R の表明集合と本稿では呼ぶ．

関係 R は外延集合と表明集合の二種の集合と紐付けられているため，これら集合は同じ関係 R で潜在的に結び付けられる．この潜在的リンクが共通空間射影法の基礎付けとなる．

2.2 関係の主-双対表現

二つ以上のエンティティが同時に出現する文は，エンティティペアの特徴付けとエンティティを結ぶ関係表現の特徴付けの二種類の観点から見ることが出来る．この二つの観点から，エンティティペアと関係表現それぞれ個別にテキスト情報から意味空間を構築出来る．エンティティペア $e^2 \in E^2$ は m 次元ベクトル $\mathbf{e}^2 \in \mathbf{E}^2 \subset \mathcal{R}^m$，関係表現 $r \in D$ は n 次元ベクトル $\mathbf{r} \in \mathbf{D} \subset \mathcal{R}^n$ で表現する．この性質から，これらの特徴付けを関係の主-双対表現と呼ぶ．

2.3 トリプレット

エンティティペアと関係表現は異なる意味空間を持っているが，テキスト中の共起情報で結び付けられる．共起情報とは，関係表現 r がテキスト中でエンティティ e_1^2 と e_2^2 を繋ぐ事を意味する．テキスト中の共起情報を頻度 $f \in \mathcal{R}$ とともに表現する記法として，トリプレット $\langle \mathbf{e}^2, \mathbf{r}, f \rangle$ を導入する．トリプレットの集合は T で定義される．この共起情報が共通空間射影法において重要な役割を占める．

2.4 共通空間射影法

二つの意味空間から共通の空間へ射影する方法として Multi-View Partial Least Squares (MVPLS) [17] を適用する．MVPLS は，Web 検索におけるクエリと文書間の類似度測定のため，クエリと文書それぞれで独自に構成した距離空間を共通空間に落として類似度算出を可能にする方法として提案された．

共通意味空間の次元数を k とする．k は $k \leq m$ および $k \leq n$ を満たす．オリジナルの意味空間から共通空間への線形射影行列を $\mathbf{L}_e$ および $\mathbf{L}_r$ と表記する．前者がエンティティペアから共通意味空間への $m \times k$ 射影行列，後者が関係表現の $n \times k$ 射影行列である．

MVPLS は直交制約を満たす線形射影行列を次の目的関数を最適化するように学習する．目的関数は，共起したエンティティペアと関係表現の共通空間におけるコサイン類似度の重み付け和で定式化される．各トリプレットは共起頻度の対数で重み付けされる．

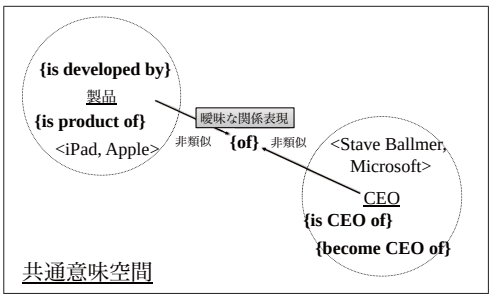

図 2: 共通意味空間における関係表現の曖昧性

$$\arg\max_{\mathbf{L}_e, \mathbf{L}_r} \sum_{(\mathbf{e}_i^2, \mathbf{r}_i, f_i) \in T} \log(f_i) \mathbf{r}_i^T \mathbf{L}_r \mathbf{L}_e^T \mathbf{e}_i^2$$
$$s.t. \quad \mathbf{L}_e^T \mathbf{L}_e = \mathbf{I}, \quad \mathbf{L}_r^T \mathbf{L}_r = \mathbf{I}\,. \tag{1}$$

得られた線形射影行列は最適化時には未使用のオブジェクトにも適用可能であり，新規オブジェクトのため最適化問題を解きなおす必要はない．この最適化問題は非凸であるが，行列 $\sum_T \log(f_i) \mathbf{e}_i^2 \mathbf{r}_i^T$ の特異値分解により，大域解を導出できる事が証明されている [17]．$\mathbf{L}_e$ は左特異行列，$\mathbf{L}_r$ は右特異行列から得られる．

2.5 共通意味空間における関係表現曖昧性

多くの関係表現は曖昧性を持つため，現実には表明集合のオブジェクトは自然にクラスタを形成しない[1]．曖昧な関係表現は複数の表明集合に所属するため，意味空間上では複数クラスタの中間に位置する事が多い．共通意味空間は共起頻度情報を反映するため，関係表現の位置は各関係 R への相対的な近接度を示すと考えられる．図 2 は共通意味空間内で曖昧な関係表現が複数クラスタの中間に位置する様子を図示している．

3 実験

本節では二つの観点から共通空間射影法の実験による評価を行う．第一に関係マイニングの観点から，共通意味空間におけるエンティティペア同士の類似度計測の有効性を量的に評価する．第二に関係表現マイニングの観点から，代表的なオブジェクトを準備し，共通空間においてそれらに意味的に類似する関係表現を抽出した時の特性を検証する．

本節では各エンティティペアを $\langle e_1, e_2 \rangle$ (例：$\langle$google, youtube$\rangle$)，関係表現を $\{r\}$ (例：{acquire}) と記述する．全タスクで ENT dataset (ENT)[2] [1] を用いる．ENT は五種類の関係カテゴリとそれらに属するエンティティペアを二十個ずつ計百個準備し，それらをクエリとした検索結果のスニペットを集めたものである．

[1] 例として，関係表現”of” は”Steve Ballmer of Microsoft” では CEO 関係，”iPad of Apple” では製品関係を表現するように，複数の関係にまたがり出現する．また，関係表現”overtake”は競争関係にある企業同士にも協力関係にある企業同士にも使われうる．

[2] http://cgi.csc.liv.ac.uk/~danushka/data/reldata.zip

	エンティティ	関係表現	トリプレット
部分単語列	12,174	12,185	521,454
最短パス	10,251	92,797	130,897

表 1: 各関係表現の定義における情報抽出結果

3.1 実験設定

3.1.1 エンティティペア・関係表現抽出

はじめにデータセットに含まれる文書を用意した．Stanford Core NLP[3]を用いて文書を文区切りにし，固有表現抽出器 [6] で文中のエンティティを抽出した．次に，二つ以上のエンティティが認識された文のみに対して，関係表現抽出を適用した．これまで関係表現は固有の定義があるものとしていたが，実際には複数の定義が考えられ，タスクに応じて定義は慎重に選択される必要がある．複雑な関係表現を扱える定義の方が一つ一つのオブジェクトがより多くの情報を扱えるため曖昧性は低くなるが，一方で出現頻度が低くなり十分なリソースがない場合ノイズの影響が強くなる．本論文では関係表現の抽出方法には詳しく立ち入らず，有効性が示されている既存の二手法を採用し，それぞれに対して性能計測を行った．関係表現が抽出されなかった文はその後の処理からは除外した．

第一の方法は，関係表現としてエンティティ間の部分単語列を列挙する方法である．深い構文解析を必要とせず，頻出パターンマイニング技術 [14] の進展もあり大規模コーパスに対しても高速で動作する．部分単語列法では，以下の 4 条件を全て満たす部分単語列を関係表現として抽出する．(1) 全単語が二つのエンティティの間に存在する．(2) 部分単語列の最大長は 6．(3) エンティティ間の単語数は 10 以下．(4) 部分単語列の出現頻度は 100 より多い．類似の抽出方法が [2] でも使用されているが，本研究は文法的制約を使用しない．結果，無意味な関係表現が大量に抽出されるが，共通空間でこれらは高い曖昧性を持つと認識される．

第二の方法は，文中の係り受け構造におけるエンティティ間の最短パスを関係表現として抽出する方法である．最短パス法は，エンティティ間の最も重要な意味は最短パスによって表現されるという仮説にもとづいている [3]．係り受け解析には Enju[4] [11, 12] を用いた．最短パスの長さは 1 から 6 に制限した．表 1 に，各関係表現の定義における抽出結果を表示する．

3.1.2 各空間の特徴ベクトル生成

エンティティペアの特徴量として，分布仮説 [8] に基づき周辺の単語を利用する．特徴量の値および特徴選択のため自己相互情報量 (PMI) を用いる．PMI は $\log_e\left(p(w_a|\langle e_i, e_j\rangle)/p(w_a)\right)$ で定義される．$p(w_a)$ は文書全体での単語 w_a の出現確率，$p(w_a|\langle e_i, e_j\rangle)$ はエンティティペア $\langle e_i, e_j\rangle$ の出現する文における単語 w_a の条件付き出現確率である．PMI の値が 1.0 以上の単語のみを特徴量として利用した．より効果的に周辺情報を利用するため，文を (1) 第一エンティティより前，(2) エンティティペア間，(3) 第二エンティティより後ろ，の三セクションに分割して，個別にスコアを計算した．ウィンドウ幅は 2 から 6 に設定し，5 または 6 で最高精度を示すことを確認した[5]．

関係表現の特徴量として，拡張分布仮説 [9] にもとづき共起したエンティティの頻度情報を用いた．第一エンティティと第二エンティティは個別に頻度を求めた．特徴量の値および特徴選択には PMI を用いた．

コンテキスト情報を基にした特徴ベクトルは高次元かつスパースになりやすいため，乱択化 SVD[7] によって次元削減を行った．次元削減は後の共通意味空間射影にて計算コストおよび使用メモリの削減に繋がる．乱択化 SVD でエンティティペアは 3000 次元，関係表現には 1000 次元の特徴ベクトルを生成した．

3.1.3 共通意味空間射影法

最後に，共通意味空間を生成するため MVPLS を適用した．共通意味空間の次元数は 1000 に設定した．

3.2 関係マイニング評価

はじめに，共通意味空間の性質を関係マイニングの観点から評価した．意味的によい計量を持つ空間は，同じ意味を持つエンティティペア同士の類似度が高くなると考えられる．既存研究 [2] に基づき，意味カテゴリ付きのエンティティペアのみに着目し，性能を評価した．各カテゴリの各エンティティペアについて，コサイン類似度の高いエンティティペア上位 10 個を抽出し，$\sum_{t=1}^{10}\mathrm{Rel}(t)\cdot\mathrm{Pre}(t)/10$ のスコアで性能を計測した．ここで，$\mathrm{Rel}(t)$ は t 位の結果が本来のエンティティペアと同じカテゴリに属していた場合 1 になる二値関数，$\mathrm{Pre}(t)$ は 1 位から t 位までの結果の適合率を返す関数である．比較のため，共通空間射影前の特徴ベクトルと関係表現との共起頻度を並べて生成した特徴ベクトルの二種類のベースラインモデルを用いた．

実験結果を表 2 に示す．この結果から，共通空間射影法で得られる特徴ベクトルはベースラインモデルに比べ，エンティティペア間の類似度を求める上でより有効であることを確認できる．この結果は既存研究 [2] の実験結果よりも高い精度を示しており，共起情報に加え特徴ベクトルの情報も用いる有効性が確認できる．

	部分単語列		最短パス	
ウィンドウ幅	5	6	5	6
Original (3,000)	0.878	0.889	0.908	**0.913**
Original (1,000)	0.878	0.881	0.905	0.911
Relation (1,000)	0.822	0.822	0.621	0.621
Embedded (1,000)	**0.887**	**0.899**	**0.912**	0.912

表 2: 関係マイニングの実験評価結果：各数字は平均評価値．各実験設定での最高精度は**太字**で示す．括弧内の数字は特徴次元数．Original：エンティティペア空間の特徴ベクトル (3.1.2 節)．Relation：関係表現との共起情報で生成した特徴ベクトル．Embedded：共通空間の特徴ベクトル (2.4 節)．

[3] http://nlp.stanford.edu/software/corenlp.shtml
[4] http://www.nactem.ac.uk/enju/

[5] これらの方法で，関係マイニングタスク (3.2 節) における精度が約 80% から約 90% に向上することを確認した．

{announce acquisition}		{president ,}	
共通意味空間	関係表現空間	共通意味空間	関係表現空間
{announce that have acquire}	{announce that have acquire}	{chairman ,}	{'s president be}
{complete acquisition}	{acquire}	{, ceo &}	{would say}
{say have it buy}	{pay}	{'s president ,}	{would that say}
{acquire}	{buy}	{ceo &}	{'s blue and}
{pay}	{compra}	{chief ,}	{'s chairman ,}
{'s acquisition}	{buy company}	{, ceo)}	{chairman ,}
{'s out_of}	{say that it buy}	{chief ,}	{palmisano}
{'s purchase}	{nor}	{would that say}	{,}
{acquisition}	{do}	{executive ,}	{, reader ,}
{'s takeover}	{announce be buy}	{ceo become}	{palmisano include door '}

表 3: 関係表現 {announce acquisition} と {president ,} に対する類似度上位 10 件の関係表現

⟨charlie chaplin, london⟩		⟨facebook inc, mark zuckerberg⟩	
共通意味空間	共起頻度順	共通意味空間	共起頻度順
{bear walworth}	{bear}	{'s executive ,}	{, ceo}
{bear april}	{'s " arrangement while lay orchestra}	{ceo be}	{, ceo (}
{play}	{,}	{ceo}	{founder and}
{bear}	{reception}	{,}	{everything , ceo}
{bear}	{'s}	{'s president ,}	{andceo}
{bear woolsthorpe ,}	{'s}	{have say}	{ceo ,}
{bear woolthrope}	{be when}	{, ceo ,}	{-}
{be member parliament}	{and}	{, ceo}	N/A
{bear woolsthorpe}	{bear april street , walworth ,}	{ceo become}	N/A
{'s}	{walk ,}	{buy}	N/A

表 4: エンティティペア ⟨charlie chaplin, london⟩ と ⟨facebook inc, mark zuckerberg⟩ に対する類似度上位 10 件の関係表現

3.3 関係表現マイニング評価

次に，関係表現に関する共通意味空間を評価する．関係表現は人手での包括的正解セットの作成が困難なため，代表的なエンティティペアと関係表現を準備し，それらに類似度の高い関係表現を集め，質的比較する．良い意味的計量を持つ空間は，意味的に近い関係表現を抽出出来るはずである．さらに，共通意味空間は曖昧な表現の自動除去が可能な事を示す．ページ数制約のため，関係表現の抽出方法は最短パス法，ウィンドウ幅が 5 の時の実験結果のみを表示する．また，出現頻度が 10 以下の最短パスは結果から除去した．

表 3 に各空間における関係表現間の類似度上位 10 個を示す．{announce acquisition} と {president ,} を代表例として用いている[6]．共通空間射影法により，{compra}，{nor}，{,} 等の曖昧な関係表現が除去され，曖昧性の低い {'s purchase} や {chief ,} 等の順位が上がっている事が確認できる．

表 4 では，出生地の関係性を持つ ⟨charlie chaplin, london⟩ と CEO の関係性を持つ ⟨facebook inc, mark zuckerberg⟩ の，共通空間における類似度上位 10 件と共起頻度上位 10 件を示す．⟨charlie chaplin, london⟩ の結果より，共起頻度順位では曖昧性の高い関係表現が多く抽出される一方，共通空間内では類似度上位から曖昧性の高いものは除去されることが確認できる．また，⟨facebook inc. mark zuckerberg⟩ の結果から，{'s executive ,} のように，直接的に共起していなくても意味的に近い関係表現は共通空間内で高い類似度を示している事が確認できる．

参考文献

[1] Danushka T. Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Measuring the similarity between implicit semantic relations from the web. In *Proc. of WWW*, 2009.

[2] Danushka T. Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Relational duality: unsupervised extraction of semantic relations between entities on the web. In *Proc. of WWW*, 2010.

[3] Razvan C. Bunescu and Raymond J. Mooney. A shortest path dependency kernel for relation extraction. In *Proc. of HLT/EMNLP*, 2005.

[4] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proc. of AAAI*, 2010.

[5] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proc. of EMNLP*, 2011.

[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of ACL*, 2005.

[7] Nathan Halko, Per G. Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[8] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[9] Dekang Lin and Patrick Pantel. Dirt - discovery of inference rules from text. In *Proc. of KDD*, 2001.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, 2013.

[11] Yusuke Miyao and Jun'ichi Tsujii. Probabilistic disambiguation models for wide-coverage hpsg parsing. In *Proc. of ACL*, 2005.

[12] Yusuke Miyao and Jun'ichi Tsujii. Feature forest models for probabilistic hpsg parsing. *Computational Linguistics*, 34(1):35–80, 2008.

[13] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. Patty: A taxonomy of relational patterns with semantic types. In *Proc. of EMNLP-CoNLL*, 2012.

[14] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440, 2004.

[15] Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proc. of EMNLP*, 2009.

[16] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL*, pages 1201–1211, 2012.

[17] Wei Wu, Hang Li, and Jun Xu. Learning query and document similarities from click-through bipartite graph with metadata. In *Proc. of WSDM*, 2013.

[6] 表 3 には {chief ,} のように表面上の単語列が同じでも内部の述語構造が異なり，異なる関係表現として扱われているものもある．