ベクトル空間モデルに基づく単語の意味表現の性質

秋山 哲史（電気通信大学大学院電気通信学研究科）
内海 彰（電気通信大学電気通信学部システム工学科）
satoshi@utm.se.uec.ac.jp, utsumi@se.uec.ac.jp

1 はじめに

様々な言語処理技術において，単語の意味をどのように扱うかは重要な問題である．単語の意味を表現する手法には様々なものがあり，その一つとして，ベクトル空間モデルに基づいた手法がある．これは，テキストデータにおける単語の出現頻度等を基に単語を多次元空間上に配置する手法であり，その多次元空間上での近さが単語同士の関係の強さを反映する．今日では，認知モデルの構築 (秋山，内海 2005) や情報検索など幅広い分野の研究に応用されている．

ベクトル空間モデルに基づいた意味表現の構築手法は，今日様々なものが提案されており (笠原 他 1997; Landure et al. 1998)，異なる手法で構築された意味表現の間には，当然ながら性質の差異が予想される．しかし，意味表現の性質の比較や，構築手法と性質の関係についての研究はほとんど行われていない．その中で，鈴木 (鈴木 2006; Utsumi, Suzuki 2006) は語義的類似と，連想関係から語義的類似に相当するものを省いた連想的類似の二つの類似性を定義することで意味表現の性質の比較を行っている．しかし，連想関係には，反意語，類義語関係といった語義的類似の他にも，統語関係や熟語関係などの文脈上での近さや,「犬」と「動物」のような概念的な上下関係など，単語間の様々な関係が包含されている．意味表現の性質を明らかにする上では，それらの関係についても区別して扱うべきである．

本研究では，単語間の様々な関係において，異なる手法で構築した意味表現を比較することにより，各意味表現の性質の差異を明らかにするとともに，手法と性質との関係を明らかにすることを目的とする．

2 意味表現の構築

ベクトル空間モデルに基づき意味表現を構築する手法の概略は以下の通りである．

1. テキストデータにおける単語の出現頻度等に基づいて，単語の特徴を表す重みを要素とする特徴ベクトルを作成し，特徴ベクトルを並べた特徴行列を構成する．
2. 構成した特徴行列において，同等の特徴が複数の次元（特徴次元）に含まれている場合がある．このような特徴を一つの特徴次元として扱うため，特徴行列の圧縮を行う．

2.1 テキストデータ

以下の２種類のテキストデータを用いる．それぞれのテキストデータを形態素解析し，名詞 · 動詞 · 形容詞を意味表現の対象とする．

新聞記事　毎日新聞 CD-ROM より，’99 年 1,4,7,10 月の 4ヶ月分を利用し，183534 段落，59431 単語を対象とする．

国語辞典　学研国語大辞典 CD-ROM を用い，75845 単語を対象とする．

2.2 特徴行列の構成

特徴行列 $\boldsymbol{M}$ は，各単語 $i(1 \leq i \leq n)$ の特徴ベクトル $\boldsymbol{w_i}$ を並べたものである．

$$\boldsymbol{M} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_i}, \cdots, \boldsymbol{w_n})^T \tag{1}$$

特徴ベクトル $\boldsymbol{w_i}$ は，単語 i における特徴 $j(1 \leq j \leq m)$ の特徴量 w_{ij} を要素とする．

$$\boldsymbol{w_i} = (w_{i1}, w_{i2}, \cdots, w_{ij}, \cdots, w_{im}) \tag{2}$$

$\boldsymbol{w_i}$ は，以下の４種類の方式に基づき決定する．

2.2.1 特徴行列の構成法式

出現頻度方式 (TF)　w_{ij} は，単語 i の文章単位 j における出現頻度とする．文章単位は一般に，文や段落等が用いられる．本研究では，新聞記事では１段落，国語辞典では見出語と語義文をあわせたものを文章単位として扱う．

共起頻度方式 (CO)　w_{ij} は，単語 i が出現する文章単位中に単語 j が出現する頻度とする．

Sliding Window 方式 (SW)　w_{ij} は，単語 j が単語 i の前後 N 単語以内に出現する頻度とする．この手法は，テキストデータ中の文章がある程度の長さを持つ必要があるが，国語辞典の語義文は非常に短いため，この方式には適さない．本研究では $N = 1, 5, 9$ の 3 パタンで特徴行列を構成する．

概念ベース方式 (CB)　笠原 他 (1997) が提案している手法．まず，w'_{ij} を単語 i の語義文における単語 j の出現頻度とする行列 $\boldsymbol{M'}$ を構成し，

$$\boldsymbol{M} = \boldsymbol{M'} + 0.2\sqrt{\boldsymbol{M'^2}} + 0.2\boldsymbol{M'^T} \tag{3}$$

となる行列 $\boldsymbol{M}$ を構成する．ここで，$\sqrt{\boldsymbol{A}}$ は行列 $\boldsymbol{A}$ の要素毎に平方根をとるものとする．$\boldsymbol{M}$ は，$\boldsymbol{M'}$ に単語 i の語義文中の単語 j の語義文に出現する単語の頻度と，単語 i を含む語義文に出現する単語の頻度を合わせたものに相当する．この方式は単語とその語義文を区別する必要があるため，新聞記事を用いることはできない．

2.3 特徴行列の次元圧縮

2.2 節で構成したそれぞれの特徴行列 $\boldsymbol{M}$ に対し，以下の３手法を用いて特徴行列の次元圧縮を行う．

2.3.1 特徴行列の次元圧縮手法

シソーラス圧縮 (TR)　概念ベース手法 (笠原 他 1997) に対して用いられている手法で，シソーラス（単語を概念の木構造で表現した階層型の辞書）上で同じ上位概念を持つ特徴次元同士を，一つの上位概念にまとめる．圧縮後の各要素は下位概念に相当する要素の単純和とする．特徴次元に単語が割り当てられている必要があるため，出現頻度方式に対して用いることはできない．

特異値分解 (SVD)　Latent Semantic Analysis(LSA)(Landure et al. 1998) で用いられている手法で，主成分分析の手法の一つである特異値分解により，特徴行列 $\boldsymbol{M}$ を

$$\boldsymbol{M} = \boldsymbol{U\Sigma V^T} \tag{4}$$

と分解し，大きい順に k 番目までの特異値を用い，

$$\hat{\boldsymbol{M}} = \boldsymbol{U_k\Sigma_k V_k^T} \tag{5}$$

と $\boldsymbol{M}$ を変換する．このときの $\boldsymbol{U_k}$ を用いて語の意味を表現する．

表 2: 連想関係の分類結果の例

刺激語	熟語	随伴	統語	全体-部分	カテゴリ-事例	類義語	反意語	事例-事例
悪魔	-	恐怖, 死	物語, 夢	-	人間, 妖精	悪	神, 善人	天使, 魔女
椅子	勉強，安楽	クッション, 学校	-	足, 木	-	-	-	机, 座布団

表 1: 連想関係の分類と例

関係	例	関係	例
熟語	満員-電車	カテゴリ-事例	動物 $\supset$ 犬
随伴	タバコ $\Rightarrow$ 煙	類義語	宣伝≒広告
統語	友人-紹介	反意語	戦争 $\Leftrightarrow$ 平和
全体-部分	手 $\supset$ 指	事例-事例	赤 $\Leftrightarrow$ 青 $\Leftrightarrow$ 緑 $\in$ 色

Random Indexing(RI)　Sahlgren (2005) が提案している手法．圧縮後の特徴次元数を k とする．$m \times k$ の乱数行列 $\boldsymbol{R_k}$ を用い，

$$\boldsymbol{M_k} = \boldsymbol{M}\boldsymbol{R_k} \tag{6}$$

と $\boldsymbol{M}$ を変換する．$\boldsymbol{R_k}$ の各要素に正規乱数を用いることで，$\boldsymbol{R_k}$ の各行ベクトルは擬似的に直行しているとみなすことができる．このときの $\boldsymbol{M_k}$ を用いて語の意味を表現する．

2.3.2　圧縮後の次元数

特異値分解は圧縮後の次元数 k を任意に指定することが出来る．本研究では $k = 200, 400, 600, 800, 1000$ として圧縮を行った．シソーラス圧縮では圧縮に用いるシソーラスにより次元数が固定される．本研究では日本語大シソーラス CD-ROM を用い，975 次元に圧縮した．

3　比較実験

2 章で示したそれぞれの手法を用いて意味表現を構築し，それらの性質について比較実験を行った．

3.1　実験方法

一般に，意味表現の性能評価を行う際には，人の連想との一致度を評価する．人の連想には単語間のさまざまな関係が混在しており，これらの関係は，ベクトル空間モデルに基づいた意味表現においては，単語同士のベクトル空間上での近さで表される．これらのことから，意味表現の性能評価には,「良い意味表現では，ある単語（刺激語）から連想されやすい単語（連想語）はベクトル空間上で刺激語の付近に配置される」という基準が用いられる．しかし，この基準では，連想の中に混在する個々の関係については区別されていないため，対象とする意味表現が実際に表現している関係については不明確である．

そこで本研究では，人の連想を複数の関係に分類し，それぞれの関係ごとに，意味表現の比較を行った．

3.1.1　連想関係の分類

文献 (阿部 他 1994) によれば，連想関係は表 1 のように分類することができる．そこで本研究では，既存の連想実験のデータを，アンケートにより表 1 に示す関係に分類した．

参加者　学生１６名

用いた連想データ　連想基準表 (梅本 1969) より選択した刺激語 60 単語と，各刺激語に対し連想された単語群（平均 24 単語），計 1440 対を対象とした．

分類方法　被験者に，刺激語と連想語のペアを表 1 とともに提示し，2 単語間の関係としてふさわしいと思われるものをすべて選択してもらった．表 1 のどの関係にも属さないものや，それ以外の関係を持っていると考えられるものについては,「その他」の項目を設け，そこに分類してもらった．各ペアあたり 6 名に回答してもらい，4 名以上が選択した関係のみを 2 単語間に存在する関係とした．分類結果の一部を表 2 に示す．

アンケートにより得られた結果をさらに次の三つの関係に大別した．

近接的関係　「熟語」,「随伴」,「統語」といった関係は，テキスト中において，近い距離で共起しやすい関係と言える．このような文脈的な近接性に依存するものを,「近接的関係」として大別した．

上位-下位関係　「全体-部分」や,「カテゴリ-事例」といった関係には概念的な上位-下位が存在する．これらを「上位-下位関係」として大別した．

語義的関係　「類義語」や「反意語」といった関係は，語の定義と密接な関わりがある．また,「事例-事例関係」は共通の語義的概念を持ったグループ内での関係であり，同時に「類義語」や「反意語」の一部を内包していると言える．これらの関係を,「語義的関係」として大別した．

表 1 に示した関係に，これら三つの関係,「その他」の関係および，全ての関係を含めた「全体」を加え，評価に用いた．

3.1.2　単語間の関係の強さ

ベクトル空間モデルでは単語間の関係の強さをそれぞれの単語に対応するベクトルの成す角の余弦やユークリッド距離などを用いて数値化することが出来る．しかし，異なる手法で構築された意味表現間では，ベクトル空間上での単語の分布の範囲が異なる．そのため，ベクトルの成す角の余弦やユークリッド距離は異なる意味表現間では共通の尺度として用いることはできない．

そこで本研究では，連想語の刺激語に対する意味表現中での関係の強さを次のようにして表した．まず，刺激語とベクトル空間上に配置された全ての単語とのなす角の余弦を求め，値の高い順に単語をソートする．このときの連想語の順位を用いて，刺激語に対する連想語の関係の強さを表した．この順位が高い連想語ほど，刺激語との関係が強いことになる．

なお，各意味表現ごとの語数の違いによる影響を排除するため，実験に用いた全ての意味表現に共通の 39405 単語を対象とした．

3.1.3　評価基準

「良い意味表現では，ある単語（刺激語）から連想されやすい単語（連想語）はベクトル空間上で刺激語の付近に配置される」という基準の具体的な表現方法としては，あ

表 3: 各関係における特徴行列の平均再現率

テキストデータ	特徴行列	**近接的関係**	熟語	随伴	統語	**上位-下位関係**	全体-部分	カテゴリ-事例	**語義的関係**	類義語	反意語	事例-事例	その他	全体
国語辞典	TF	**0.53**	0.64	0.48	0.62	**0.68**	0.59	0.84	**0.77**	0.69	0.84	0.70	0.30	0.54
	CO	**0.50**	0.42	0.52	0.56	**0.58**	0.55	0.64	**0.49**	0.43	0.51	0.66	0.51	0.51
	CB	**0.13**	0.12	0.15	0.11	**0.20**	0.14	0.26	**0.33**	0.26	0.40	0.26	0.05	0.16
新聞記事	TF	**0.60**	0.62	0.58	0.61	**0.53**	0.45	0.65	**0.56**	0.55	0.56	0.64	0.40	0.55
	CO	**0.52**	0.52	0.50	0.56	**0.48**	0.44	0.53	**0.49**	0.42	0.51	0.51	0.45	0.50
	SW1	**0.23**	0.22	0.22	0.27	**0.26**	0.22	0.30	**0.35**	0.32	0.36	0.38	0.14	0.23
	SW5	**0.47**	0.51	0.44	0.54	**0.42**	0.38	0.48	**0.49**	0.42	0.52	0.48	0.37	0.45
	SW9	**0.49**	0.52	0.47	0.53	**0.44**	0.39	0.51	**0.50**	0.44	0.53	0.46	0.39	0.47

る刺激語に対する連想語を正解とし，意味表現中での刺激語に対する連想語の近さの順位が n 位以上のものを出力とした際の再現率 R_n や適合率 P_n が考えられる．

$$\text{再現率 } R_n = \frac{\text{上位 } n \text{ 位以内の連想語数}}{\text{各関係における連想語の総数}} \quad (7)$$

$$\text{適合率 } P_n = \frac{\text{上位 } n \text{ 位以内の連想語数}}{n} \quad (8)$$

本研究ではこの基準をそれぞれの関係ごとに用いて評価を行うことを目的とするが，3.1.1 節で分類したデータはそれぞれの関係において，連想語の総数に偏りがあるので，同一手法における各関係の優劣を比較する際には，連想語の総数に依存しない評価基準が必要となる．上位 n 位以内に連想語が含まれる確率は連想語の総数が多いほど高くなる．このため，P_n は連想語の総数に依存してしまうが，R_n は連想語の総数を分母とするので，この条件を満たすことができる．また，上位 n 位以内の連想語数が等しい場合であっても，連想語がより上位に分布する場合と n 位付近に分布する場合とを比較した場合，前者の方がより良い意味表現であるといえる．そこで，本研究では，前述した条件を満たし，かつこのような比較が可能な評価基準として，n を値を変化させて R_n を計算し平均をとる平均再現率 R_{ave} を定義し，関係ごとに算出して評価に用いた．なお，R_n の値は総単語数の 10% を基準に上位 4000 位まで 100 位刻みで算出した．

$$\text{平均再現率 } R_{ave} = Ave(R_{100}, R_{200}, ..., R_{4000}) \quad (9)$$

この値が高いほど，それぞれの関係を表現するうえで有効な手法といえる．

3.2 実験１：特徴行列の性質の評価

まずはじめに，次元圧縮を行う前の特徴行列の性質を調査した．対象とした特徴行列は表 3 に示した 8 種類である．

3.2.1 結果

各特徴行列の各関係における R_{ave} の値を表 3 に示す．表中の太字は大別した関係，下線はそれぞれの関係において R_{ave} の最も高かったものを示している．

近接的関係　「熟語」,「統語」では国語辞典・TF が,「随伴」では新聞記事・TF が最も高い値を示し,「近接的関係」全体としては新聞記事・TF が最も高い値を示した．また，国語辞典よりも新聞記事で高い値となる傾向が見られた．

上位-下位関係　「全体-部分」,「カテゴリ-事例」ともに国語辞典・TF が最も高い値を示した．また，新聞記事よりも国語辞典で高い値となる傾向が見られた．

語義的関係　「類義語」,「反意語」,「事例-事例」の全てにおいて，国語辞典・TF が最も高い値を示した．

表 4: TF，CO の大別した各関係と「全体」の R_{ave} の差

テキストデータ	特徴行列	近接的関係	上位-下位関係	語義的関係
国語辞典	TF	-0.01	0.14	0.23
	CO	-0.01	0.07	-0.01
新聞記事	TF	0.05	-0.02	0.01
	CO	0.02	-0.02	-0.01

全体　新聞記事・TF が最も高い値を示した．CO よりも TF で高い値を示す傾向がみられ，SW では N が大きくなるにつれて R_{ave} が向上する傾向が見られた．また，国語辞典・CB は他の手法に比べ極端に低い値となった．

3.2.2 考察

テキストデータの性質　TF，CO どちらの手法においても,「近接的関係」では新聞記事を用いた場合に,「語義的関係」および「上位-下位関係」においては国語辞典を用いた場合により高い値を示す傾向が見られた．国語辞典は語義を簡潔に記述しているため,「上位-下位関係」や,「語義的関係」のように厳密な定義の可能な関係を良く表現できたと考えられる．反面，語義文中の各文には文脈的な繋がりがほとんど無いため,「近接的関係」はうまく表現できなかったと考えられる．一方，新聞記事では各文が文脈的な繋がりを持っているため,「近接的関係」を良く表現することができたと考えられる．

出現頻度と共起頻度の性質　表 4 は TF および CO における，大別した各関係と「全体」との R_{ave} の差を示したものである．国語辞典では「上位-下位関係」,「語義的関係」を，新聞記事では「近接的関係」を良く表現できることがわかったが，この傾向は CO よりも TF でより強く表れていることがわかる．テキストにおける単語の出現頻度は共起頻度に比べ，より表層的な情報であるといえる．CO に比べ TF でテキストごとの性質の変化が大きくなったのはテキストにおける表層的な情報のみを捉えたためと考えられる．一方，単語の共起頻度はテキストに内在する，より複雑な情報といえる．CO では捉えた情報が複雑であるがゆえに，TF ほど明確な性質が見られなかったと思われる．全体としてで TF に比べ CO の値が低くなったのも，共起頻度の情報が複雑であるために，語の意味を表現する上でのノイズとなったとためと考えられる．

3.3 実験２：次元圧縮の効果の評価

3.2 節で用いた各特徴行列に対して 2.3 節で示した各手法を用いて次元圧縮を行い，次元圧縮による性質の変化を調査した．

表 5: 大別した関係における TR による次元圧縮の効果

テキストデータ	特徴行列	近接的関係	上位-下位関係	語義的関係	全体
国語辞典	CO	$\mathbf{0.54(0.03)}$	$\mathbf{0.73(0.15)}$	$\mathbf{0.68(0.19)}$	$\mathbf{0.58(0.08)}$
	CB	$\mathbf{0.35(0.22)}$	$\mathbf{0.52(0.32)}$	$\mathbf{0.75(0.42)}$	$\mathbf{0.39(0.23)}$
新聞記事	CO	$\mathbf{0.53(0.01)}$	$\mathbf{0.51(0.03)}$	$\mathbf{0.57(0.08)}$	$\mathbf{0.51(0.01)}$
	SW1	$\mathbf{0.36(0.13)}$	$\mathbf{0.52(0.26)}$	$\mathbf{0.47(0.12)}$	$\mathbf{0.36(0.13)}$
	SW5	$\mathbf{0.52(0.05)}$	$\mathbf{0.52(0.10)}$	$\mathbf{0.59(0.10)}$	$\mathbf{0.49(0.04)}$
	SW9	$\mathbf{0.53(0.04)}$	$\mathbf{0.52(0.08)}$	$\mathbf{0.59(0.20)}$	$\mathbf{0.51(0.04)}$

表 6: 大別した関係における SVD による次元圧縮の効果（一部）

テキストデータ	特徴行列	次元数	近接的関係	上位-下位関係	語義的関係	全体
国語辞典	CO	200	$0.41(-0.09)$	$0.48(-0.10)$	$\mathbf{0.51(0.19)}$	$0.41(-0.10)$
新聞記事	TF	200	$0.36(-0.24)$	$0.34(-0.19)$	$0.40(-0.16)$	$0.32(-0.23)$

表 7: 大別した関係における RI による次元圧縮の効果（一部）

テキストデータ	特徴行列	次元数	近接的関係	上位-下位関係	語義的関係	全体
国語辞典	TF	200	$0.15(-0.38)$	$0.23(-0.45)$	$0.30(-0.47)$	$0.17(-0.37)$
	CO	200	$0.48(-0.02)$	$0.57(-0.01)$	$0.47(-0.02)$	$0.49(-0.02)$
	CB	200	$\mathbf{0.13(0.00)}$	$0.16(-0.04)$	$0.27(-0.06)$	$0.15(-0.01)$
新聞記事	TF	200	$0.13(-0.47)$	$0.15(-0.38)$	$0.13(-0.43)$	$0.09(-0.21)$
	CO	200	$0.51(-0.01)$	$0.45(-0.03)$	$0.47(-0.02)$	$0.44(-0.01)$

3.3.1 結果と考察

大別した関係における TR，SVD および RI による次元圧縮後の R_{ave} の値を表 5, 表 6, 表 7 に示す．表中の括弧内の値は圧縮前との差を，太字は値が向上したもの示している．

TR 大別した関係における TR による次元圧縮の結果を表 5 に示す．この手法を用いることの出来る全ての特徴行列において，全ての関係で R_{ave} の向上が見られた．特に，国語辞典を用いた手法において,「上位-下位関係」および「語義的関係」の R_{ave} が大きく向上した．

特に国語辞典・CB における変化は顕著で,「語義的関係」においては圧縮前の特徴行列において R_{ave} の最も高かった国語辞典・TF とほぼ同等の値となった．一方で，それ以外の関係では国語辞典・TF に比べ大きく下回っていることから，国語辞典・CB・TR の組合せは「語義的関係」に最も特化した手法と言える．

また，国語辞典・CO では,「上位-下位関係」において，圧縮前の特徴行列を含めた全手法中で最も高い値となった．

SVD 大別した関係における SVD による次元圧縮の結果の一部を表 6 に示す．国語辞典・CO では「語義的関係」のみで R_{ave} の向上が見られた．また，圧縮前の特徴行列において R_{ave} の最も高かった新聞記事・TF は次元圧縮により R_{ave} は極端に低下する結果となった．

このほか，表 6 には示していないが，新聞記事・SW，国語辞典・TF および新聞記事・CO についてもほぼ全ての関係において R_{ave} が低下する結果となった．また，国語辞典・CB では全ての関係で R_{ave} の向上が見られたものの，どの関係においても国語辞典・CO を上回るような結果は得られなかった．

RI 大別した関係における RI による次元圧縮の結果の一部を表 7 に示す．国語辞典・CO，国語辞典・CB，新聞記事・CO および新聞記事・SW では R_{ave} はほとんど変化しなかった．一方，国語辞典・TF，新聞記事・TF では R_{ave} の著しい低下が見られ，圧縮後の次元数が低いほどその程度は強くなった．このことから，RI は特徴次元を単語で表現した特徴行列において，性能を低下させずに次元数を減らすことができることがわかった．

3.4 まとめ

「近接的関係」は新聞記事・TF,「上位-下位関係」は国語辞典・CO・TR,「語義的関係」は国語辞典・TF の組合せでそれぞれ最も良く表現できることがわかった．

また，国語辞典・CB・TR は「語義的関係」のみを表現する際に最も有効であることがわかった．

4 おわりに

今日提案されている様々な意味表現の構築手法に対し，単語間の様々な関係において比較を行うことにより，それぞれの手法の特異性や，それぞれの関係を表現する上で適切な意味表現の構築手法を見出すことができた．今後はこれらの結果をもとに，特定の関係のみを強調，あるいは排除して意味表現を構築する手法を模索していきたい．

参考文献

秋山 哲史， 内海 彰. (2005). 語の意味空間を用いた隠喩の理解モデル. 日本認知科学会 22 回大会, 257–266.

阿部 純一, 桃内 佳雄, 金子 康朗, 李 光五. (1994). 人間の言語情報処理. サイエンス社,

笠原 要, 松澤 和光, 石川 勉. (1997). 人間の言語情報処理. 情報処理学会論文誌, *38*(7), 1272–1283.

Landauer, T. K., Foltz, P. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2&3), 259–284.

Sahlgren, M. (2005). An introduction to random indexing. *Proc. of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*.

鈴木 大介. (2006). ベクトル空間モデルにおける単語の意味表現と自動要約への応用. 平成１７年度 システム工学専攻博士前期課程修士論文.

梅本 堯夫. (1969). 連想基準表－大学生 1000 人の自由連想による－. 東京大学出版会.

Utsumi, A., & Suzuki, D. (2006). Word vectors and two kinds of similarity. *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*., 858–865.