word2vecに基づく述語項構造の分布表現獲得

岩井　美樹
愛媛大学工学部情報工学科
miki@ai.cs.ehime-u.ac.jp

二宮　崇
愛媛大学大学院理工学研究科電子情報工学専攻
ninomiya@cs.ehime-u.ac.jp

1　はじめに

本研究は word2vec の手法に基づき述語項構造に対する分布表現を獲得する手法を提案する。word2vec は単語に対する概念を分布表現 (低次元の密なベクトル) として獲得する手法の一つであり、ベクトルの計算により意味のコンポジショナリティーを実現する性質を持つことから、近年注目を集めている。word2vec は、大量のテキストを用い、単語予測の擬似的なタスクをニューラルネットワークで学習することによって、単語に対する分布表現を獲得する。本研究は、述語項構造付き英語 HPSG の構文解析を行うことによって述語項構造が付与されたテキストを生成し、これらのテキストからニューラルネットワークを用いて動詞と目的語のペアに対する分布表現を獲得する。

本研究は、word2vec の手法に基づき述語項構造の分布表現を獲得する 4 つの手法を提案し、異なる述語項構造の表現が同じ意味表現となることを実験により検証する。述語項構造とは、ある文の中で述語が他の単語とどのような関係にあるのかを記述した構造である。以下の 2 つの例文は述語項構造は異なるが同じような意味表現を持つ例を示している。

1. I wash the dishes.
2. I do the dishes.

1. の述語項構造は “wash(dishes)”, 2. の述語項構造は “do(dishes)” と、2 つの例文の述語項構造は異なる構造となっている。しかし、意味に注目するとどちらも “皿を洗う” という意味を表していることがわかる。このように述語構造は異なるが同じような意味表現を持つ英文は多く存在する。本研究では、“do dishes” のようにその動詞の意味が明示的にわからない動詞-目的語のペア (‘do’ と ‘dishes’ のペア) に対し、最もその概念に近い基本動詞 (‘wash’) や述語項構造 (“wash(dishes)”) を正解として、分布表現の距離を測ることにより、各提案手法の性能を検証する。

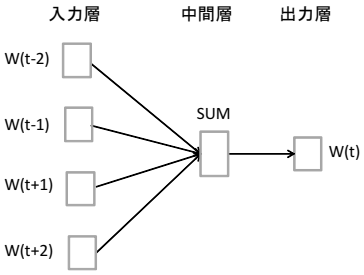

図 1: CBOW モデルの仕組み

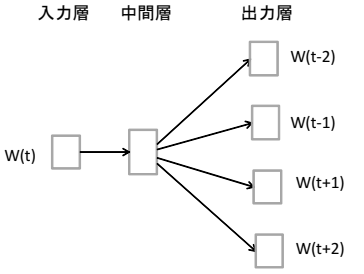

図 2: skip-gram モデルの仕組み

2　word2vec

word2vec[1] は Tomas Mikolov らによって提案されたニューラルネットワークを用いた単語分布表現 (語彙の概念を表す低次元の密なベクトル) の獲得手法である。テキスト中の各単語を周辺の単語から予測する擬似的な単語予測のタスクを設定し、このタスクを大量のテキストからニューラルネットワークで学習し、中間層における各単語の重みを抽出することによって、単語に対する分布表現を獲得する。

Mikolov らは word2vec を実現するネットワーク構造として、CBOW モデルと Skip-gram モデルの二つのモデルを提案している。図 1 は CBOW モデルのネットワーク構造を表している。図が表すように CBOW モデルは入力層、中間層、出力層からなり、周辺単語 $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_k$ を入力とし、w_t を出力とする (予測する) ニューラルネットワークである。入

力層と出力層は辞書中の単語と一対一対応のノードから成っており、1-of-K 単語ベクトルとなっている。中間層は隠れ変数となっており、任意の個数のノードを与えることができる。入力層と中間層をつなぐ重み行列は各単語に対する重みベクトルを並べたものとなっている。単語に対する重みベクトルの次元数は中間層のノード数と等しく、word2vec において数百程度がよく用いられる。つまり、入力層における語彙数 (数十万) の次元を数百程度に次元削減していることになる。CBOW モデルでは、周辺の各単語 $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_k$ に対する重みベクトルの和を計算し、それを中間ノードの値としている。

図 2 は Skip-gram モデルのネットワーク構造を表している。Skip-gram モデルでは、ある単語 w_t を入力とし、周辺の単語 $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_k$ をそれぞれ出力とする (予測する) モデルとなっている。CBOW モデルでは周辺の単語に対する重みベクトルの和を計算していたが、Skip-gram モデルでは、単語 w_t を入力としてと周辺の単語を一つ一つ予測するモデルとなっている。

CBOW モデルでも Skip-gram モデルでも、入力層と中間層をつなぐ重み行列 (各単語に対する重みベクトルの集合) が、word2vec が最終的に生成する単語分布表現となる。単語を分布表現に変換することにより単語を意味空間上の 1 点に対応させることができ、意味的に関連の強い単語グループに分類したり、単語に対する意味的な計算を可能とする利点がある。

3 提案手法

本研究で提案する述語項構造に対する分布表現の獲得手法について説明する。本研究では 4 つの手法を提案する。手法 2、手法 3、手法 4 では、述語項構造付き英語 HPSG の構文解析を行い、述語項構造付きテキストから分布表現の獲得を行う。実験では構文解析に英語構文解析器の「Enju2.4.1」を利用した。

3.1 手法 1

手法 1 では述語項構造解析を行わず word2vec をそのまま用いて、テキストから単語に対する分布表現を獲得する。次に得られた単語に対する分布表現を用いて述語項構造に対する分布表現を作成する。動詞 v の単語分布表現を $\mathbf{x}$、目的語 o の単語分布表現を $\mathbf{y}$ としたとき、$\mathbf{x}+\mathbf{y}$ を v と o に対する述語項構造の分布表現とする。

3.2 手法 2

手法 2 は word2vec の CBOW モデルにおいて、周辺単語または出力に動詞が含まれるとき、その目的語も周辺単語に含める手法である。まず、訓練コーパスに対し述語項構造付き英語 HPSG の構文解析を行う。次に、訓練コーパスのテキスト中に出現する各動詞に対し、その目的語をその動詞の直後に挿入する。具体的には、Enju の解析結果から「verb_arg12(動詞-目的語)」の関係を出力した箇所のみ、「動詞 + 空白文字 + 目的語」に変更し、下の例のようなテキストを新たに作成する。

- 変更前： My mother read Shakepeare to me. He has agreed to have lunch with me.
- 変更後: my mother **read shakespeare** shakespeare to me. he have **agree have** to **have lunch** lunch with me.

新しく作成されたテキストを word2vec に学習させ、述語項構造に対する分布表現は手法 1 と同様に、動詞の単語分布表現を $\mathbf{x}$、目的語の単語分布表現を $\mathbf{y}$ としたとき $\mathbf{x}+\mathbf{y}$ を述語項構造に対する分布表現とする。

3.3 手法 3

手法 3 は、動詞とその目的語を一つの単語とみなして CBOW モデルの学習を行う手法である。手法 2 と同様にまず訓練コーパスに対し述語項構造付き英語 HPSG の構文解析を行う。次に、訓練コーパスのテキスト中に出現する各動詞 v に対し、v とその目的語 o を連結した「$v:o$」を一つの単語として、v を $v:o$ に置き換える。以下に例を示す。

- 変更前： My mother read Shakespeare to me. He has agreed to have lunch with me.
- 変更後: my mother **read:shakespeare** shakespeare to me. he have **agree:have** to **have:lunch** lunch with me.

続いて、v を $v:o$ に置き換えた訓練コーパスを用いて CBOW モデルの学習を行い、単語に対する分布表現と $v:o$ に対する分布表現を獲得する。手法 3 においては、$v:o$ に対する分布表現が動詞 v と目的語 o の述語項構造に対する分布表現となる。

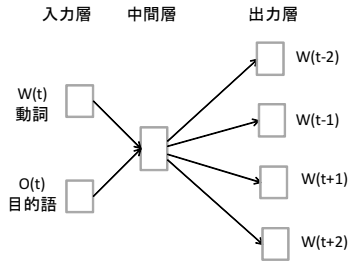

図 3: 手法 4 のニューラルネットワークの仕組み

3.4 手法 4

skip-gram モデルはある単語 w_t を入力とし、その周辺の単語を予測するモデルであるが、手法 4 は skip-gram モデルを改良し、ある単語 w_t が動詞であったとき、w_t とその目的語 w_o を入力とし、その周辺の単語を予測するモデルとしてニューラルネットワークの学習を行う。手法 4 のニューラルネットワークの構造は、図 3 のようになり、skip-gram モデルと異なり、入力は二つの単語 (動詞とその目的語) となる。中間層では動詞 w_t に対する重みベクトルと目的語 w_o に対する重みベクトルの和を計算し、それを中間ノードの値としている。述語項構造に対する分布表現は手法 1,2 と同様に、動詞の単語分布表現を $\mathbf{x}$、目的語の単語分布表現を $\mathbf{y}$ としたとき $\mathbf{x}+\mathbf{y}$ を述語項構造に対する分布表現とする。

4 実験

4.1 実験設定

本研究では word2vec を用いて実験を行った。word2vec にはいくつかのオプションがあり、学習のモデルについては、手法 1,2,3 では CBOW モデルを用い (-cbow1)、手法 4 では skip-gram モデルを改良したモデルを用いた。word2vec のウィンドウサイズは最大 8 単語 (-windows 8) とし、ネガティブサンプリングのネガティブサンプルの個数は 25 個 (-negative 25) とし、階層的ソフトマックスは使用しなかった (-hs 0)。また、中間層のノード数は 200 次元とした。

ニューラルネットワークを学習するための訓練コーパスとして、英語大規模コーパスである English Gigaword 4th edition(LDC2009T13, nyt_eng, 199412,199504, 約 1055 万語) と Corpus of Contemporary American English(COCA), Corpus of Historical AmericanEnglish(COHA) を用いた。COCA とＣＯＨＡは合わせて約 14 万語から成る。

表 1: 述語項構造に対する分布表現の候補

述語項構造	正解とする意味表現	例文
do-dish	wash	I do the dishes.
do-nail	paint, put, dress	I do my nails.
do-cleaning	clean,wash	I do the cleaning.
have-lunch	eat	I have a lunch.
have-tea	drink	I have a tea.
make-call	call	I make a call.
make-bed	clean, put, set	I make the bed.
finish-coffee	drink	He finished his coffee.
read-shakespeare	read	I read Shakespeare.
enjoy-movie	watch,see	You enojy the movie.

表 1 は、述語の意味が明示的には表れていない動詞-目的語のペアとその本来の意味を表す基本動詞の例を 10 個 [1] あげている。各手法の評価は、これらの基本動詞を正解として考え、動詞-目的語ペアに対する分布表現との距離を測ることによって行った。基本動詞の候補集合は、英語基本動詞活用辞典 [2] に収録されている 385 個の基本動詞とした。

4.2 実験結果

表 2 は提案手法である手法 1 から 4 までのそれぞれの結果を表している。表中の「順位」は、表 1 に記載される動詞-目的語ペアと基本動詞 385 個とのコサイン類似度をそれぞれ計算し、正解例となる基本動詞が 385 個中何位だったかを示している。また、ランキングの精度を測るために平均逆順位 (Mean Reciprocal Rank, MRR) の評価を行った。

表 2 より、今回のデータセットでは手法 3 が最も良い平均順位と平均 MRR スコアとなった。また、手法 2 は手法 1 よりも良い平均順位と平均 MRR となっているため、動詞とその目的語をデータに組み込むことでより良い述語項構造に対する分布表現が得られていると考えられる。一方、word2vec の入力層を 2 入力に変更し、skip-gram モデルを利用した手法 4 は全手法の中で最も低い順位と MRR スコアとなってしまったが、この原因の解明については将来の課題とする。

得られた述語項構造に対する分布表現の性質を分析するため、“read-shakespeare” と “make-bed” の 2 つの述語項構造に対して次の実験を行った。表 2 から

[1] 動詞-目的語のペアの候補の選択には「たった 3 つの動詞で日常生活の 8 割を表現する方法」という web ページ (http://english-leaders.com/hot-three-verbs/(2015 年 1 月 20 日参照)) を参考にした。

表 2: 実験結果

	述語項構造	正解とする意味表現	手法 1		手法 2		手法 3		手法 4	
			順位	MRR	順位	MRR	順位	MRR	順位	MRR
1	do-dish	wash	35	0.0286	12	0.0833	1	**1**	15	0.0667
2	do-nail	paint	65	0.0024	71	0.0019	54	**0.3812**	10	0.0571
		put	20		31		8		80	
		dress	176		107		1		17	
3	do-cleaning	clean,	21	0.0284	7	0.0839	4	**0.2083**	3	0.1783
		wash	110		40		6		43	
4	have-lunch	eat	1	**1**	**1**	1	5	0.2	3	0.3333
5	have-tea	drink	2	0.5	1	**1**	1	**1**	2	0.5
6	make-call	call	1	**1**	1	**1**	1	**1**	1	**1**
7	make-bed	clean	67	0.0195	27	0.0449	22	0.0028	13	**0.0452**
		put	25		11		33		203	
		set	276		148		103		116	
8	finish-coffee	drink	2	0.5	2	0.5	1	**1**	2	0.5
9	read-shakespeare	read	1	**1**	1	**1**	1	**1**	1	**1**
10	enjoy-movie	watch	10	0.1056	10	0.1333	113	0.0023	2	**0.375**
		see	9		6		27		4	
平均			51.3125	0.4206	29.75	0.4864	**23.8125**	**0.5841**	32.1875	0.4056

“read-shakespeare” は “read” と非常に近いことがわかるが、本来の「本を読む」という意味と近くなっているかどうかわからない。そこで、“read-shakespeare” に対し、“read-book”、“read-people” と “feel” に対するコサイン類似度を測ってみた。表 3 はその結果を表している。“read-shakespeare” は “read”,“read-book” など本に関する単語はコサイン類似度が近く、“read-people” とは離れていることがわかる。“read-people” という述語項構造は “He can’t read people.” という例文のように雰囲気を読むという意味で使用されることが多い表現であるため、“read-shakespeare” の意味表現と差異化できていることがわかる。

表 3: read-(shakespeare, people) とのコサイン類似度

	read-shakespeare とのコサイン類似度	read-people とのコサイン類似度
read-people	0.2542	-
read-shakespeare	-	0.2542
read	0.4718	0.2207
read-book	0.2849	0.2706
feel	0.2009	0.3170

表 4: make-bed とのコサイン類似度

	make-bed とのコサイン類似度
go-bed	0.5471
sleep	0.2976
clean	0.1407

表 4 は “make-bed” と “go-bed” を比較した結果を示している。“make-bed” は “Please, make the bed.” のようにベッドを整える、整頓するという意味を持つにもかかわらず、“go-bed” や “sleep” と近くなってしまい、この場合には述語項構造を概念表現できていないことがわかる。

5 おわりに

本研究は word2vec の手法に基づき述語項構造に対する分布表現を獲得する手法を提案した。述語項構造付き英語 HPSG の構文解析を行うことによって述語項構造が付与されたテキストを生成し、生成されたテキストに対し単語予測の擬似的なタスクをニューラルネットワークで学習することによって、述語項構造に対する分布表現を獲得した。提案した手法の中では、動詞と目的語のペアを一つの単語とみなして学習する手法が最も良い平均順位と平均 MRR スコアを与えた。将来の課題として、評価のための述語項構造の例を増やすこと、良い分布表現が得られない場合の原因を解明すること、良い分布表現を得るためのモデルの改良があげられる。

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (2013) “Efficient Estimation of Word Representations in Vector Space”.

[2] 小西　友七 (1980) 『英語基本動詞活用辞典』 研究社.