線型部分空間に基づく学習済み単語埋込空間上の集合演算

石橋 陽一 [1]　横井 祥 [2,3]　須藤克仁 [1,3,4]　中村哲 [1,3]
[1] 奈良先端科学技術大学院大学　[2] 東北大学　[3] 理研 AIP　[4] 科学技術振興機構さきがけ
{ishibashi.yoichi.ir3, sudoh, s-nakamura}@is.naist.jp　yokoi@tohoku.ac.jp

概要

単語埋込は自然言語処理の基盤的な道具であり単語単位の表現学習は進んでいる。一方で加法構成に代表される意味計算は、単語を集合として扱っていながら集合の演算までは定義できていない。これを現代の単語埋込空間の枠組みで表現できれば集合と埋込表現の両方の特性を反映した優れた表現を作ることができる。そこで本研究では事前学習済み単語埋込空間で集合・集合演算を教師なしで表現することを目指した。我々は線型部分空間に基づく集合演算である量子論理に着目し、埋込空間で線型部分空間が言語的な集合演算として機能することを実証し、文の意味的類似性タスクへ応用した。

1　はじめに

単語埋込は現代の自然言語処理の基盤技術でありGloVe [1] や word2vec [2] 等の静的表現や、BERT [3] 等の動的表現が、あらゆる種類の言語処理タスクの性能を大きく押し上げた [4]。

単語埋込は単語に対して与えられる表現だが、実際の処理対象は単語**の集合**であることが多い。例えば、単語間の意味の階層構造を見出すことができる [5]。例えば red, blue, green, . . . からなる集合は Color というクラスを表していると捉えられる。こうした概念集合に対する計算は自然言語処理のために重要である [6]。また、句、文、文書など単語より大きな単位の計算において最も基本的で強力なアプローチは、これらを単語集合とみなすことである。例えば文の類似性を計算するタスク（STS） [7] における基本的な指針では、文の類似性を単語集合の重複度（意味の重複度）に帰着させる [8]。

このように単語集合の計算には強いニーズがある。単語をシンボルで表現していた時代には単語集合を形式的な集合として扱っていたが [9]、単語をベクトルで表現する場合の単語 “集合” の扱いは近似的なものに留まっている。例えば句や文の表現として単語ベクトルの和を用いる加法構成 [10] は強力であるが、単語 “集合” としての性質がどのように入っているかは不明である。また、単語集合の重複率を計算するためにこれを適合率・再現率 [11] や最適輸送コスト [12, 13] 帰着させる向きもあるが、“集合” の計算としては近似的なアプローチに過ぎない。

そこで本研究では、現代の単語埋込の枠組みで集合演算を再度定式化することを試み、事前学習済み単語埋込空間上で集合および集合演算が線型部分空間で表現できることを実証した。

2　集合を扱うために必要な演算

本研究では和集合や共通部分など集合を扱う各種演算を、単語埋込空間で近似的にかつ単語埋込のリッチな情報を活用しつつ教師なしで計算することを目指す。具体的には、単語埋込空間上で帰属関係

$$\mathrm{Color} = \{\mathrm{red}, \mathrm{blue}, \ldots\},\ \mathrm{Fruit} = \{\mathrm{apple}, \mathrm{peach}, \ldots\} \quad (1)$$

$$\mathrm{orange} \in \mathrm{Color} \cap \mathrm{Fruit} \quad (2)$$

や集合間類似度（Jaccard 係数）

$$\mathrm{A} = \{\mathrm{A}, \mathrm{boy}, \mathrm{walk}, \ldots\},\ \mathrm{B} = \{\mathrm{The}, \mathrm{child}, \mathrm{run}, \ldots\} \quad (3)$$

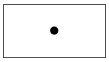
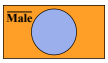
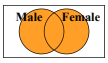
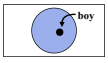
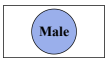
$$\mathrm{Jaccard}(\mathrm{A}, \mathrm{B}) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \quad (4)$$

を計算可能にすることを目指す。これらの計算を実現するために必要な演算は以下の通り（表 1 左）。

- **元と集合の表現**：元となる単語とそれをまとめた対象である集合を埋込空間で表現する。
- **集合に対する演算**：ひとつまたは複数の集合に対して、補集合・和集合・共通部分を計算する。
- **帰属関係の計算**：元が集合に属しているかどうかを判定する。
- **集合の濃度の計算**：集合に含まれる元の数（濃度）を計算する（式 (4)）。

次のセクションでこれらの表現や計算を単語埋込空間で近似的に実現する方法を提案する。

表 1　集合と埋込空間での表現の対応

集合		埋込空間での表現
元 king		ベクトル $\boldsymbol{v}_{king}$
集合 Male = {king, man, . . .}		部分空間 $\mathbb{S}_{\text{Male}}$ = span$\{\boldsymbol{v}_{king}, \boldsymbol{v}_{man}, \ldots\}$
補集合 $\overline{\text{Male}}$		直交補空間 $(\mathbb{S}_{\text{Male}})^{\perp}$
和集合 Male ∪ Female		和空間 $\mathbb{S}_{\text{Male}} + \mathbb{S}_{\text{Female}}$
共通部分 Color ∩ Fruit		共通部分 $\mathbb{S}_{\text{Color}} \cap \mathbb{S}_{\text{Fruit}}$
帰属関係 boy ∈ Male		帰属度 Member$(\mathbb{S}_{\text{Male}}, \boldsymbol{v}_{boy})$
濃度 |Male|		部分空間の次元 dim $\mathbb{S}_{\text{Male}}$

3　線型部分空間による集合表現

2 節で示した集合演算を埋込空間で実現する方法を述べる。集合演算として必要な性質を保証しつつ単語埋込空間の幾何的な特性を活用するため、本研究では量子論理を単語埋込空間へ適用する。

3.1　量子論理

量子論理とは量子力学的現象を表現する理論の 1 つである [14]。粒子などの状態を元とした集合の処理という要請に対して、量子論理は集合を**線型部分空間**の言葉で記述し、さらにド・モルガンの法則、二重否定などの成立を理論的に保証している。

3.2　単語埋込空間での集合演算

量子論理はヒルベルト空間の上で記述されるが、単語埋込空間（ユークリッド空間）もヒルベルト空間でありほとんどの演算はそのままユークリッド空間での表現に翻訳することができる。不足している演算（帰属の度合い、濃度）に関しては量子論理と一貫した方法を提案する。提案法のまとめは表 1 の通り。このあと詳細について述べる。

元と集合の表現　単語集合 A を $\{\mathrm{w}_1, \mathrm{w}_2, \ldots\}$、単語 w に対応する単語ベクトルを $\boldsymbol{v}_w$ で表す。ここで、我々がまず表現したい対象は元と集合であった。本研究では、元である単語 w を単語ベクトル $\boldsymbol{v}_w$ で、単語集合 $\mathrm{A} = \{\mathrm{w}_1, \mathrm{w}_2, \ldots\}$ を単語ベクトルが**張る線型部分空間**で表現する。

$$\mathbb{S}_{\mathrm{A}} = \mathbb{S}_{\{\mathrm{w}_1, \mathrm{w}_2, \ldots\}} := \mathrm{span}\{\boldsymbol{v}_{w_1}, \boldsymbol{v}_{w_2}, \ldots\} \tag{5}$$

以降、線型部分空間を単に「部分空間」と呼ぶ。

集合に対する演算　ある集合 A の補集合 $\overline{\mathrm{A}}$ は、部分空間 $\mathbb{S}_{\mathrm{A}}$ の**直交補空間** $(\mathbb{S}_{\mathrm{A}})^{\perp}$ で表現する。

$$\mathbb{S}_{\overline{\mathrm{A}}} = (\mathbb{S}_{\mathrm{A}})^{\perp} := \{\boldsymbol{u} \mid \exists \boldsymbol{v} \in \mathbb{S}_{\mathrm{A}}, \boldsymbol{u} \cdot \boldsymbol{v} = 0\} \tag{6}$$

ある集合 A と B の和集合 $\mathrm{A} \cup \mathrm{B}$ は、部分空間 $\mathbb{S}_{\mathrm{A}}$ と $\mathbb{S}_{\mathrm{B}}$ の**和空間** $\mathbb{S}_{\mathrm{A}} + \mathbb{S}_{\mathrm{B}}$ で表現する。

$$\mathbb{S}_{\mathrm{A} \cup \mathrm{B}} = \mathbb{S}_{\mathrm{A}} + \mathbb{S}_{\mathrm{B}} := \{\boldsymbol{u} + \boldsymbol{v} \mid \boldsymbol{u} \in \mathbb{S}_{\mathrm{A}}, \boldsymbol{v} \in \mathbb{S}_{\mathrm{B}}\} \tag{7}$$

ある集合 A と B の共通部分 $\mathrm{A} \cap \mathrm{B}$ は、部分空間 $\mathbb{S}_{\mathrm{A}}$ と $\mathbb{S}_{\mathrm{B}}$ の**共通部分（交空間）** $\mathbb{S}_{\mathrm{A}} \cap \mathbb{S}_{\mathrm{B}}$ で表現する。

$$\mathbb{S}_{\mathrm{A} \cap \mathrm{B}} = \mathbb{S}_{\mathrm{A}} \cap \mathbb{S}_{\mathrm{B}} = \{\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{S}_{\mathrm{A}}, \boldsymbol{v} \in \mathbb{S}_{\mathrm{B}}\} \tag{8}$$

帰属関係の計算　最も簡単な方法として、単語の集合に対する帰属関係（例: boy ∈ Male）は、ベクトルの部分空間に対する**帰属関係**（例: $\boldsymbol{v}_{boy} \in \mathbb{S}_{\text{Male}}$）で表現できる。ただし一見自然なこの方法はベクトル空間の持つ「近さ」の概念を活用しきれない。例えば Male 集合に含まれていない単語 nephew に関して、$\boldsymbol{v}_{nephew}$ が $\mathbb{S}_{\text{Male}}$ と非常に近い位置に存在していたとしても「帰属関係はない」と 2 値で判断されてしまう。そこで本研究では、ベクトル $\boldsymbol{v}_w$ と部分空間 $\mathbb{S}_{\mathrm{A}}$ の近さに応じて連続値を返す**帰属度**の関数 Member$(\mathbb{S}_{\mathrm{A}}, \boldsymbol{v}_w)$ を定義する。

$$\theta_{\mathbb{S}_{\mathrm{A}}, \boldsymbol{v}_w} = \min \left\{ \arccos \left(\frac{|\boldsymbol{u} \cdot \boldsymbol{v}_w|}{\|\boldsymbol{u}\| \|\boldsymbol{v}_w\|} \right) \middle| \, \boldsymbol{u} \in \mathbb{S} \right\} \tag{9}$$

$$\mathrm{Member}(\mathbb{S}_{\mathrm{A}}, \boldsymbol{v}_w) = \cos \theta_{\mathbb{S}_{\mathrm{A}}, \boldsymbol{v}_w} \in [0, 1] \tag{10}$$

$\theta_{\mathbb{S}_{\mathrm{A}}, \boldsymbol{v}_w}$ は $\boldsymbol{v}_w$ と $\mathbb{S}_{\mathrm{A}}$ の「なす角」（第一正準角）で、単語埋込空間で単語ベクトル同士のなす角が意味的類似度を表現することを間接的に利用することができる。$\boldsymbol{v}_w \in \mathbb{S}_{\mathrm{A}}$ の場合帰属度は 1 で、$\boldsymbol{v}_w$ が $\mathbb{S}_{\mathrm{A}}$ と直交する場合は 0 である。

集合の濃度の計算　ある単語集合 A の濃度（単語数）を部分空間 $\mathbb{S}_{\mathrm{A}}$ の**次元** dim $\mathbb{S}_{\mathrm{A}}$ で表現する。

3.3　Subspace Jacccard

和集合・共通部分・濃度に関する部分空間での計算方法を用いれば、Jaccard 係数（式 (4)）を部分空間で計算できる。すなわち、単語ベクトル集合に対する重複度を自然に計算することができる。この新しい尺度を SubspaceJacccard と呼ぶ。

$$\mathrm{SubspaceJacccard}(\mathbb{S}_{\mathrm{A}}, \mathbb{S}_{\mathrm{B}}) = \frac{\dim \mathbb{S}_{\mathrm{A}} \cap \mathbb{S}_{\mathrm{B}}}{\dim \mathbb{S}_{\mathrm{A}} + \mathbb{S}_{\mathrm{B}}} \in [0, 1] \tag{11}$$

表 2 単語集合データセットの統計情報。「集合」は Male = {king, man, . . .} のような集合演算を適用していない集合を表す。#正例、#負例は集合あたりの平均要素数。

単語集合	集合の数	#正例	#負例
集合	11	44.2	73.5
補集合	11	934.8	3.5
和集合	110	88.3	699.9
共通部分	8	1.5	55.9

4 埋込空間上の集合演算の実証

提案法で作った部分空間 $\mathbb{S}_{\text{Male}}$ は、単語集合 Male = {king, man, . . .} 全体が表す「男性的」という意味を良く表現できているだろうか？これを検証するため定量的・定性的両面から実験を行った。

4.1 実験設定

この実験で共通する設定について述べる。

単語埋込 Common Crawl (840B tokens) で事前学習済みの 300 次元 GloVe [1] と Google News で事前学習済みの 300 次元 word2vec [2] を用いた。

データセット 検証のため、本研究ではある単語集合に対して帰属関係が成立する単語（正例）$\mathcal{P}$ と、成立しない単語（負例）$\mathcal{N}$ のデータを作成した（表 2）。正例 $\mathcal{P}$ のデータ (S, w) は集合 S、集合の元である単語 $\mathrm{w} \in \mathrm{S}$ からなり（例 $(\text{Male}, \text{boy}) \in \mathcal{P}$）、負例 $\mathcal{N}$ のデータ $(\mathrm{S}, \mathrm{w}')$ は集合と集合の元でない単語 $\mathrm{w}' \notin \mathrm{S}$ からなる（例 $(\text{Male}, \text{apple}) \in \mathcal{N}$）。作成方法の詳細は付録 A に記載する。

4.2 部分空間外のベクトルの評価

提案法が単語集合を良く表現するなら、帰属関係 boy ∈ Male や apple ∉ Male が提案法でも成立するはずである。これ検証するため次の実験を行った。

検証方法 Male 集合を例に説明する。まず Male 集合から単語 boy を取り除いた空間 $\mathbb{S}_{\text{Male}\setminus\{\text{boy}\}}$ を作る。次にこの空間が、取り除いた boy を意味的に含んでいるかを帰属度 $\text{Member}(\mathbb{S}_{\text{Male}\setminus\{\text{boy}\}}, \boldsymbol{v}_{boy})$ で確認する。この値が Male に含まれていない単語、例えば apple に対する帰属度 $\text{Member}(\mathbb{S}_{\text{Male}\setminus\{\text{boy}\}}, \boldsymbol{v}_{apple})$ よりも高ければ、提案法の Male 集合の表現 $\mathbb{S}_{\text{Male}}$ は Male の意味を良く捉えていると言える。同様の実験を補集合・和集合・共通部分でも行う。

1） https://nlp.stanford.edu/projects/glove/
2） https://code.google.com/archive/p/word2vec/

表 3 帰属度の平均。括弧内は標本標準偏差。

	正例 ↑	負例 ↓	Δ ↑
集合	0.84 (± 0.14)	0.58 (± 0.17)	0.26
補集合	0.79 (± 0.16)	0.55 (± 0.15)	0.24
和集合	0.88 (± 0.10)	0.70 (± 0.13)	0.18
共通部分	0.66 (± 0.14)	0.14 (± 0.09)	0.52

例えば $\text{Member}(\mathbb{S}_{\text{Color}\setminus\{\text{orange}\}} \cap \mathbb{S}_{\text{Fruit}\setminus\{\text{orange}\}}, \boldsymbol{v}_{orange})$ が $\text{Member}(\mathbb{S}_{\text{Color}\setminus\{\text{orange}\}} \cap \mathbb{S}_{\text{Fruit}\setminus\{\text{orange}\}}, \boldsymbol{v}_{rice})$ より高いことを確かめる。

評価尺度 正例 $(\mathrm{S}, \mathrm{w}) \in \mathcal{P}$ の集合 S の部分空間 $\mathbb{S}$ に $\boldsymbol{v}_w$ が帰属し、負例 $(\mathrm{S}, \mathrm{w}') \in \mathcal{N}$ では $\mathbb{S}$ に $\boldsymbol{v}_{w'}$ が帰属しないことを、平均帰属度の差 Δ で評価する。

$$\frac{1}{|\mathcal{P}|}\sum_{(\mathrm{S},\mathrm{w})\in\mathcal{P}} \text{Member}(\mathbb{S}, \boldsymbol{v}_w) - \frac{1}{|\mathcal{N}|}\sum_{(\mathrm{S},\mathrm{w}')\in\mathcal{N}} \text{Member}(\mathbb{S}, \boldsymbol{v}_{w'})$$

これが正の値であれば部分空間で boy ∈ Male や apple ∉ Male が表現できていることが示唆される。

実験結果 結果（表 3）から、「正例との帰属度」「負例との帰属度」に差があり、量子論理が単語集合・集合演算を表現できていることがわかった。

4.3 部分空間内のベクトルの評価

提案法が単語集合を良く表現できているなら「男性的」集合表現 $\mathbb{S}_{\text{Male}}$ の中には「男性的」なベクトル $\boldsymbol{v}_{boy}$ が含まれているはずである。そこで、提案法で作った部分空間内に含まれているベクトル（単語ベクトルとは限らない）の最近傍単語を確認する。

検証方法 この実験では正例の単語集合のみ使用する（表 2）。ある単語集合の単語の 50%を使用して部分空間を張り、それ以外の単語を検証用とした。ここでは、Color ∩ Fruit 集合を例にとり説明する。まず部分空間 $\mathbb{S}_{\text{Color}} \cap \mathbb{S}_{\text{Fruit}}$ に含まれるベクトル $\boldsymbol{v}$ を無作為にサンプリングする。単語埋込の語彙全体の単語ベクトル集合に対して $\boldsymbol{v}$ の最近傍単語ベクトル $\boldsymbol{v}_w$ を取得し、最近傍単語 w が Color ∩ Fruit の元らしいかを確認する。例えば最近傍単語が orange で部分空間を張る単語でないならば、色と果物の集合の共通部分が良く表現できていると言える。この検証をいくつかの集合におこなった。

実験結果 結果は表 4 の通り。word2vec の結果については付録 B を参照されたい。表中の「部分空間表現の近傍単語」は部分空間を張る際に使用していない単語のみ列挙している。この結果は単語集合の言語的特徴が部分空間に良く反映できていることを

表 4 単語集合の部分空間内のベクトルと cos 類似度上位 3 件の単語。使用した単語集合は Male = {his, male, . . .}、Female = {queen, heroine, . . .}、Color = {purple, white, . . .}、Fruit = {lime, grape, . . .}。

表現する集合	部分空間表現	部分空間の近傍単語 (GloVe)	部分空間の近傍単語 (word2vec)
Male	$\mathbb{S}_{\text{Male}}$	brother-in-law, nephews, stepfather	father, grandkids, son
$\overline{\text{Male}}$	$(\mathbb{S}_{\text{Male}})^{\perp}$	drunken, erect, nominal	Ahnlund, Confirm, Sakowicz
Male ∪ Female	$\mathbb{S}_{\text{Male}} + \mathbb{S}_{\text{Female}}$	moms, uncle, boy	siblings, bachelors, giantess
Color ∩ Fruit	$\mathbb{S}_{\text{Color}} \cap \mathbb{S}_{\text{Fruit}}$	orange, peach, mango	pear_apple, pear, orange

表 5 各 STS タスクの相関係数

		STS12	STS13	STS14	STS15	STS16
Skip-Thought		41	29	40	46	52
GloVe	Avg-cos	52.1	49.6	54.6	56.1	51.4
	DynaMax	58.2	53.9	65.1	70.9	71.1
	BERTScore	52.8	47.2	62.1	67.3	-
	Subspace	52.7	50.9	59.7	67.7	61.2
BERT	Avg-cos	47.8	53.5	58.3	64.0	64.6
	DynaMax	51.2	49.1	58.6	68.4	65.0
	Subspace	49.5	43.9	55.9	67.1	63.9

示している。例えば、$\mathbb{S}_{\text{Male}}{}^{\perp}$ 内のベクトルは Male 以外の単語、$\mathbb{S}_{\text{Color}} \cap \mathbb{S}_{\text{Fruit}}$ 内のベクトルは色と果物に共通する orange と類似しており、$\mathbb{S}_{\text{Male}} + \mathbb{S}_{\text{Female}}$ 内のベクトルは男女両方の特徴を持っている。

5 実験: 教師なし文類似度タスク

単語埋込空間上の集合演算を用いた集合間類似度を教師なし文類似度タスク (STS) に応用し、埋込空間の集合間類似度として機能するかを確かめる。

検証方法 STS では 2 つの文（単語集合）の類似度を算出し人手評価との相関係数で評価する。

実験設定 データセットは SemEval shared task の 2012-2016 [7, 15, 16, 17, 18] のものを使用した。ベースラインは平均ベクトルの cosine 類似度（Avg-cos）と [8] の DynaMaxJaccard [3)]、Skip-Thought Vector [19] を使用した。評価尺度にはピアソンの相関係数を使用した。単語埋込は 4 節と同じ GloVe、word2vec、そして BERT-base [3] の最終層のトークンに対応する動的埋込を使用した。なお教師なし STS のためファインチューニングはおこなっていない。

実験結果 結果は表 5 の通り。埋込空間での傾向は BERT よりも GloVe で提案法の効果が高かった word2vec での結果も GloVe と同様な傾向ため付録 B に記載する。GloVe では提案法が全てのタスクで加法構成に基づく手法（Avg-cos）を上回った。量子論理によって集合と集合演算を部分空間で表現したことで、静的埋込において集合間類似度を単語埋込空間上の演算へ拡張できたことを示している。

3) ベースラインのスコアは [8, 12] の論文から参照した。

6 関連研究

ここでは埋込空間で集合の表現・演算を試みている先行研究を概観し、提案法との違いを述べる。

教師なし集合表現 [8] は Fuzzy 集合の考えに基づき単語埋込空間上で集合および集合演算を Fuzzy Bag of Words という方法で表現している。BERTScore [11] は集合間の適合率と再現率を内積行列を使って近似的にモデル化している。Word Rotator's Distance [12] は単語ベクトル集合間の最適輸送に基づき類似度をモデル化した。これらの研究は擬似的な集合表現、もしくは一部の集合演算の実現のみに留まっている。それに対し提案法は線型部分空間に基づく各集合演算の表現と集合に関する法則の成立が理論的に保証されている強みがある。

教師あり集合表現 DeepSet [6] は集合データに対する教師あり表現学習の代表的な手法である。本研究では事前学習済み埋込空間上で教師なしで単語集合・集合演算を表現した。

7 結論

本研究では、単語埋込の枠組みによる集合演算を線型部分空間を用いて定式化した手法を提案し、単語埋込空間において集合・集合演算を実現した。

謝辞

本研究は JST さきがけ (JPMJPR1856)、JST ACT-X（JPMJAX200S）、奈良先端科学技術大学院大学支援財団支援事業の支援を受けたものです。

参考文献

[1] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL**, pp. 1532–1543, 2014.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In **Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.**, pp. 3111–3119, 2013.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[4] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.

[5] George A. Miller. WordNet: A Lexical Database for English. **Commun. ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[6] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, **Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA**, pp. 3391–3401, 2017.

[7] Eneko Agirre, Daniel M. Cer, Mona T. Diab, and Aitor Gonzalez-Agirre. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, **Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012**, pp. 385–393. The Association for Computer Linguistics, 2012.

[8] Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Francesco Moramarco, Jack Flann, and Nils Y. Hammerla. Don't Settle for Average, Go for the Max: Fuzzy Sets and Max-Pooled Word Vectors. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.

[9] Christopher D. Manning and Hinrich Schütze. **Foundations of statistical natural language processing**. MIT Press, 2001.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019**, pp. 3980–3990. Association for Computational Linguistics, 2019.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[12] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word Rotator's Distance. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020**, pp. 2944–2960. Association for Computational Linguistics, 2020.

[13] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From Word Embeddings To Document Distances. In Francis R. Bach and David M. Blei, editors, **Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015**, Vol. 37 of **JMLR Workshop and Conference Proceedings**, pp. 957–966. JMLR.org, 2015.

[14] Garrett Birkhoff and John Von Neumann. The logic of quantum mechanics. **Annals of mathematics**, pp. 823–843, 1936.

[15] Eneko Agirre, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic Textual Similarity. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, **Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA**, pp. 32–43. Association for Computational Linguistics, 2013.

[16] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In Preslav Nakov and Torsten Zesch, editors, **Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014**, pp. 81–91. The Association for Computer Linguistics, 2014.

[17] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, **Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015**, pp. 252–263. The Association for Computer Linguistics, 2015.

[18] Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, **Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016**, pp. 497–511. The Association for Computer Linguistics, 2016.

[19] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-Thought Vectors. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, **Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada**, pp. 3294–3302, 2015.

A　単語集合データの構築方法

単語集合データ 2 の構築方法について説明する。正例セットは WordNet [5] の名詞の上位語（例：“Animal”）を単語集合のクラス、下位語（例：“dog”）を集合の元として収集し、下位語の数が 20 以上のものを使用した。これらの単語集合をランダムに組み合わせ補集合・和集合・共通部分を作成した。次に、負例を作成した。英単語の頻度データ [4] を使用し頻度上位 3000 件からランダムに 1000 件をサンプリングし、正例の単語を除去して負例セットを構築した。

B　word2vec での実験結果

表 6　帰属度の平均。括弧内は標本標準偏差。

	正例 ↑	**負例** ↓	Δ ↑
集合	0.77 (± 0.16)	0.53 (± 0.17)	0.24
補集合	0.76 (± 0.15)	0.52 (± 0.16)	0.24
和集合	0.82 (± 0.13)	0.67 (± 0.14)	0.16
共通部分	0.46 (± 0.11)	0.12 (± 0.002)	0.34

表 7　各 STS タスクの相関係数

		STS12	**STS13**	**STS14**	**STS15**	**STS16**
word2vec	Avg-cos	51.6	58.2	65.6	67.5	64.7
	DynaMax	53.7	59.5	68.0	74.2	71.3
	BERTScore	47.8	43.5	56.3	62.1	-
	Subspace	50.7	55.1	60.9	68.3	61.1

4）https://github.com/PrincetonML/SIF/