人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第21回)
SIG-AM-21-11

空間座標情報を用いた単語の分散表現獲得

Learning Word Embeddings Using Spatial Information

城光英彰　岡隆之介　内出隼人　伍井啓恭　大塚貴弘

Hideaki Joko, Ryunosuke Oka, Hayato Ucihde, Hiroyasu Itsui and Takahiro Ohtsuka

三菱電機株式会社　情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation

Abstract: This study proposes a word embedding learning method for performing both facility and region name search. In the previous method, the word embeddings were learned based on the context in which the words emerged. However, because the previous method missed the spatial information, the word embeddings that share similar contexts (e.g., universities) become similar even though the objects of the words were different (e.g.,“東京大学”(The University of Tokyo) and“京都大学”(Kyoto University)). In the proposed method, the word embeddings are learned based on the spatial information data, which comprise both the object names and coordinates. Therefore, even if the words share similar contexts, the proposed method can learn different word embeddings for different words if those objects are different. The proposed method is evaluated using the synonym search task. As a result, it was observed that the MRR for the evaluation data improved using the proposed method in comparison with the previous method. Furthermore, the proposed method improved the Mean Reciprocal Rank by 177% maximally, i.e., from 0.151 to 0.418, in comparison with the previous method.

1. はじめに

施設名や地名の検索などにおいて，単語の文字面だけでなく「意味」を考慮した検索が求められている.「意味」を考慮した検索には，分散表現の有効性が知られている[1] [2] ．一般に，この分散表現は，「類似した意味の単語は，その文脈の単語（文脈単語）の分布も類似する」という分布仮説[3] に基づき獲得される．しかし，既存手法では，空間の情報を考慮しないため，異なる対象を表す分散表現が類似する問題がある[4] ．例えば，「東京大学」と「京都大学」は対象としては異なるが，どちらも「大学」という共通の特徴を持つために，その単語の周辺文脈には似たような単語が出現するため，二つの単語は類似した分散表現となる．その結果，例えば，東京大学への行き方を検索しようと「東大」(東京大学の略称）をクエリ単語として入力したときに，京都大学への行き方が検索されてしまう問題が生じうる.

この問題を解決するために，本研究では，空間座標情報を用いた分散表現獲得手法を提案する．空間座標の情報は個別の対象ごとに異なると想定されるため, この情報を活用すれば異なる対象については，相違した分散表現が得られると考えられる．提案手法は，自然言語で記述された名称とその空間座標情報からなる大規模データ（空間座標 DB）から分散表現を学習する．本研究では，空間座標 DB として，ウィキペディア日本語版から取得した地名・施設名とその空間座標からなるデータを用いた.

評価は同義語検索タスクにより行った．具体的には，クエリ単語（例えば「東大」）と各検索対象単語（例えば，「東京大学」や「京都大学」）について，分散表現のコサイン類似度を算出し，これに基づき検索対象単語をランキングし，そのランキング精度を平均逆順位（MRR）により評価した.

2. 既存手法（Skip-gram モデル）

分布仮説のもとで分散表現を獲得する研究には，Mikolov et al.[1] の Skip-gram モデルがある．本節では，Skip-gram モデルについて，先行研究[2] を参考に概説する. Skip-gram モデルは分布仮説に基づき単語の意味を表す数値ベクトル（分散表現）を獲得する手法である. ある単語w_tが文章内の位置tに存在した場合の，その文脈単語$w_{t+j}(j \neq 0)$の発生確率$p(w_{t+j}|w_t)$を以下の式で与える.

$$p(w_{t+j}|w_t) = \frac{e^{v'(w_{t+j})^T v(w_t)}}{\sum_w e^{v'(w)^T v(w_t)}}$$

ここで，$v(w_t)$は単語w_tの分散表現，$v'(w_{t+j})$は文脈単語の出現確率計算用のベクトルである．学習はテキストデータ内の全単語に対し行われる．そのため，尤度目的関数l_{SG}は，以下の式で定義される．

$$l_{SG} = \sum_{t=1}^{T} \sum_{-c \le j \le c,\ j \ne 0} \log p(w_{t+j}|w_t)$$

ここでTはテキストデータのサイズ，cは文脈窓のサイズである．学習時は，尤度目的関数lを最大化する分散表現$v(w)$を求める．図 1 に分布仮説を用いた分散表現獲得手法のイメージを示す．中心単語「東大」の分散表現は，中心単語の文脈の単語分布により計算される．しかし，本手法では，原理的に，文脈単語の分布が類似する単語同士の分散表現は類似する．そのため，例えば「東京大学，京都大学」などの対象としては異なるが，文脈単語の分布が類似するものは，類似した分散表現となり問題である．

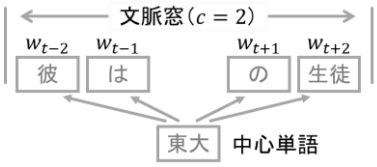

図 1: 分布仮説を用いた分散表現獲得手法のイメージ図

3. 提案手法

3.1 Skip-space モデル

既存手法の問題を解決するために，本研究では，空間座標情報を用いた分散表現獲得手法（Skip-space モデル）を提案する．提案手法は，自然言語で記述された名称とその空間座標情報からなる大規模データである空間座標 DB から分散表現を学習する．空間座標 DB のイメージを図 2 に示す．単語「東大」を含む対象の近傍に出現する対象の名称に含まれている単語（近傍単語）と，単語「東京大学」の近傍単語はともに「文京」「本郷」などを含んでおり，類似している．一方で，「東京大学」と「京都大学」などの対象として異なるものは，空間座標 DB での近傍単語の分布が相違するため，相違した分散表現になる．その結果，第 1 節に例示した，東京大学への行き方を検索しようと「東大」をクエリ単語として入力した場合に，京都大学への行き方が検索されてしまう問題を解決できる．

以下では，提案手法の近傍単語の発生確率および尤度目的関数について述べる．対象$x_c \in X$に対し，x_cからユークリッド距離が近い順に対象をk個取得することを考える．取得する対象を近傍対象$x_s \in S_{x_c}$と呼ぶ．対象x_cおよび近傍対象x_sの名称の先頭からt番目およびu番目の単語を$w_{x_c,t}$および$w_{x_s,u}$とする（図 3 を参照）．このとき，提案手法 Skip-space モデルでは，近傍単語の発生確率および尤度目的関数l_{SS}を次のように変更する．

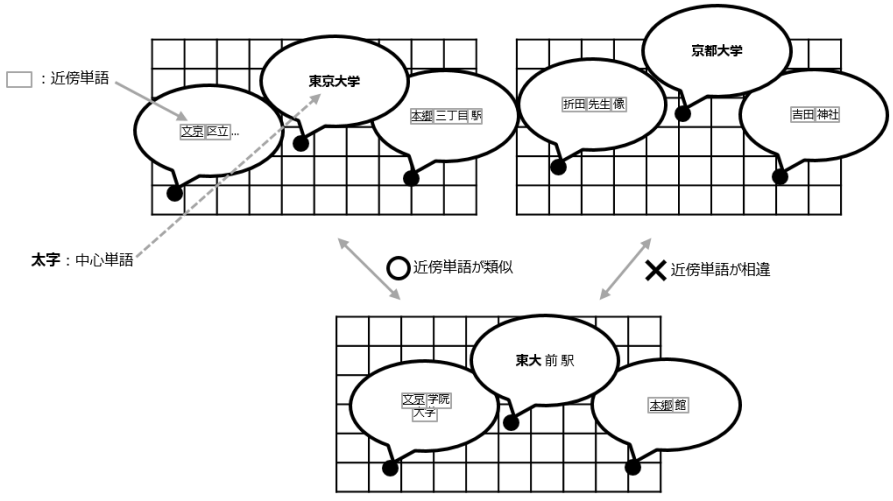

図 2: 空間座標 DB のイメージ図．単語「東大」を含む対象の近傍に出現する対象の単語（近傍単語），と，単語「東京大学」の近傍単語は類似する．

$$p(w_{x_s,u}|w_{x_c,t}) = \frac{e^{v'(w_{x_s,u})^T v(w_{x_c,t})}}{\sum_w e^{v'(w)^T v(w_{x_c,t})}}$$

$$l_{SS} = \sum_{x_c \in X} \sum_{x_s \in S_{x_c}} \sum_{1 \le t \le N_{x_c}} \sum_{1 \le u \le N_{x_s}} \frac{1}{N_{x_s}} \log p(w_{x_s,u}|w_{x_c,t})$$

ここで，N_{x_c}およびN_{x_s}は対象x_cおよび近傍対象x_sの名称を構成する単語の数である．また，$\frac{1}{N_{x_s}}$は各単語の近傍単語の出現数の偏り[1]をなくすための正規化用の係数である．

3.2 Multi-task Skip-space モデル

Multi-task Skip-space モデルでは，空間座標データとテキストデータからマルチタスク学習[5] により分散表現を学習する．尤度目的関数l_{SS+SG}が次のように変更される．

$$l_{SS+SG} = l_{SS} + l_{SG}$$

ここで，l_{SS}およびl_{SG}は，Skip-space および Skip-gram の尤度目的関数である．なお，文脈単語のベクトルv'は Skip-space による空間座標データ学習時と，Skip-gram によるテキストデータ学習時で，個別に与えられる．すなわち，各単語wは二つの文脈単語のベクトル$v_{sg}'(w)$, $v_{ss}'(w)$を保持し，文脈単語および近傍単語の発生確率は，Skip-gram によるテキストデータ学習時においては，

$$p(w_{t+j}|w_t) = \frac{e^{v_{sg}'(w_{t+j})^T v(w_t)}}{\sum_w e^{v_{sg}'(w)^T v(w_t)}}$$

Skip-space による空間座標データ学習時においては，

$$p(w_{x_s,u}|w_{x_c,t}) = \frac{e^{v_{ss}'(w_{x_s,u})^T v(w_{x_c,t})}}{\sum_w e^{v_{ss}'(w)^T v(w_{x_c,t})}}$$

のように，変更される．

[1] 対象ごとにその名称を構成する単語の数は異なるため，近傍単語の出現数の偏りが生じる．

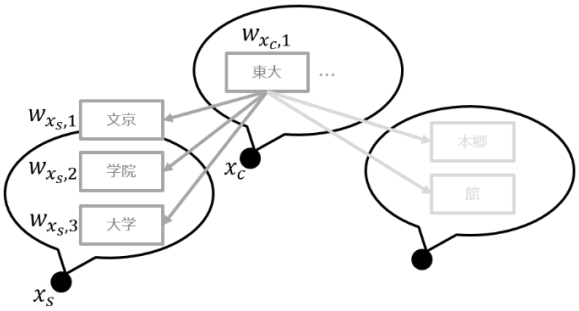

図 3: 空間座標 DB のイメージ図．単語「東大」を含む対象の近傍に出現する対象の単語（近傍単語），単語「東京大学」の近傍単語は類似する．

4. 関連研究

単語の分散表現を獲得する研究には様々なものがある．城光ら[2] は Skip-gram の拡張として，文脈限定 Skip-gram を提案した．文脈限定 Skip-gram は，文脈単語を特定の品詞を持つものや特定の位置に存在するものに限定し，各限定条件に対して分散表現を学習することで，単純な Skip-gram モデルに比べて高い精度で同義語獲得ができる．この手法は，品詞が同義語獲得に与える影響の分析が可能なため，解釈可能性が高く，さらに，限定条件の変更も容易であり，拡張可能性も高いという利点がある．しかし，この手法は単純な Skip-gram と同様に，この手法は空間の情報を考慮しないため，異なる対象について，相違した分散表現を得ることは難しい．

Skip-gram とは異なるアプローチで単語の分散表現を獲得する研究もある．Peters ら[6] は，Deep Bi-directional Language Model の隠れ層を加重平均することで，文脈を考慮した分散表現を獲得できるモデルである Embeddings from Language Models (ELMo)を提案した．Devlin [7] らは，Bi-directional Transformer [8] を用いて分散表現を獲得する Deep Bidirectional Transformers (BERT)を提案した．これらのモデルは，質問応答を含む様々な自然言語処理タスクにおいて state-of-the-art な性能を達成しており，用いるモデルも Skip-gram モデルより複雑なことから，異なる対象について，相違した分散表現を得ることができる可能性はある．しかし，これらのモデルで獲得した単語の分散表現を利用するには，都度その周辺文脈から分散表現を再計算する必要がある．再計算に必要な計算機の性能を考慮すると，機器への組み込みなどによる実用化は難しい．

これらの研究と異なり，提案手法は空間座標の情

報を活用できるため，異なる対象については相違した分散表現が得られるという優位点がある．また，単語の分散表現をその周辺文脈から再計算する必要がないため低算量であるという優位点もある．

5. 実験

実験では，まず，提案手法および既存手法により分散表現の学習を行う．用いた分散表現学習用データおよび手法については 4.1 節に示す．

評価は同義語検索タスクにより行う．具体的には，入力クエリと各検索対象の地名・施設名のそれぞれの分散表現からコサイン類似度を算出し，コサイン類似度の高い順に検索対象単語をランキングし，そのランキング精度を MRR により評価する．この評価実験のイメージを図 4 に示す．なお，地名・施設名が複数の形態素から構成される場合（例えば「都留 文化 大」など）は，各形態素の分散表現の平均値を名称のベクトルとする[9] [6] ．形態素解析は MeCab[2]を用いて行い，形態素辞書には IPAdic[10] を用いた．

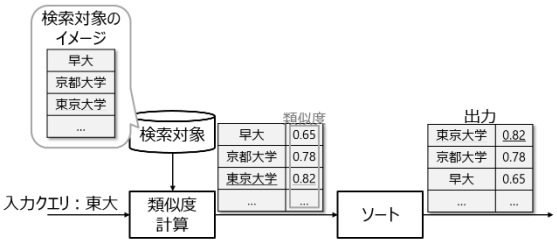

図 4: 評価実験のイメージ図．入力クエリと各検索対象の地名・施設名について，分散表現のコサイン類似度を算出し，これに基づき検索対象単語をランキングし，そのランキング精度を MRR により評価する

4.1 分散表現学習用データ

用いた分散表現学習用データについて，表 1 に示す．「地名・施設名」を緯度・経度情報が付与されている記事の名称と定義する．提案手法 Skip-space モデルにおける学習用データは，ウィキペディア日本語版から取得した地名・施設名とその空間座標データからなる「地名・施設名 空間座標データ」である．既存手法における学習用データは，DBpedia Japanese から取得した地名・施設名と，その “dbpedia-owl:abstract” プロパティから取得した概要テキストからなる「地名・施設名　概要テキストデータ」である．これらのデータに対し，第 5 節に示したのと同様の方法で形態素解析した．なお，地名・施設名概要テキストデータの取得は，このデータの述べ形態素数が，地名・施設名 空間座標データの延べ形態素数を超えた時点で終了し，提案手法と既存手法で用いたデータの延べ形態素数が揃うようにした．

表 1: 使用した分散表現学習用データ

分散表現学習用データの名称	種別	入手元	施設数	延べ形態素数
地名・施設名 空間座標データ	空間座標データ	ウィキペディア 日本語版	122K	456K
地名・施設名 概要テキストデータ	テキストデータ	DBpedia Japanese	—	456K

4.2 分散表現学習手法

まず，各分散表現学習手法と，各々の手法に対し使用した学習データを表 2 に示す．SS とは提案手法である Skip-space モデルを「地名・施設名 空間座標データ」に対し適用したものである．SS+SG とは，提案手法である Multi-task Skip-space モデルを「地名・施設名 空間座標データ」と「地名・施設名 概要テキストデータ」に対し適用したものである．SG とは，既存手法である Skip-gram モデルを「地名・施設名概要テキストデータ」に対し適用したものである．なお，学習データの詳細については 4.1 節を参照されたい．

次に，実験で使用したハイパーパラメータの設定を述べる．分散表現の次元数は，既存手法，提案手法ともに 200 とした．また，既存手法の文脈窓のサイズ$c = 5$，提案手法において取得する近傍の対象数$k = 5$とした．なお，学習の際には高速化のために Hierarchical softmax [11] による近似を行った．

[2] http://taku910.github.io/mecab/

表 2: 各分散表現学習手法と，各々の手法に対し使用した学習データ

分散表現学習手法の名称	使用した分散表現獲得モデル	使用した分散表現学習用データ	
		地名・施設名 空間座標データ	地名・施設名概要テキストデータ
提案手法: SS	Skip-space モデル	○使用	×未使用
提案手法: SS+SG	Multi-task Skip-space モデル	○使用	○使用
既存手法: SG	Skip-gram モデル	×未使用	○使用

4.3 評価用データ

用いた評価用データは，ウィキペディア日本語版から取得した「大学の略称データ」, DBpedia Japanese から取得した「DBpedia 地名・施設名 略称データ」および「DBpedia 地名・施設名 別名データ」の三種類である．SS，SS+SG, SG の少なくとも一つの手法で分散表現が獲得されていない単語を含む地名・施設名をもつ評価データは，評価用データから除外した．評価用データの詳細を表 3 に示す．また，評価用データの例を表 4 に示す．各地名・施設名は意味 ID を付与されており，異なる地名・施設名が同一の意味 ID を持つ場合，その地名・施設名は同一の意味を持つ語（同義語）であることを示す．例えば，表 4 では，「東京大学」と「東大」がどちらも“1”という同一の意味 ID を持つ．そのため，「東京大学」と「東大」は同義語であることがわかる．

評価の際には，評価用データの中から一つを入力クエリとして，残りを検索対象として使用する．そのため，「大学の略称データ」においては，各クエリについて検索対象は 259 件，「DBpedia 地名・施設名略称データ」では 62 件，「DBpedia 地名・施設名 別名データ」では，186 件となる．

表 3: 評価用データの詳細

名称	入手元	データ数
大学の略称データ	ウィキペディア 日本語版	260
DBpedia 地名・施設名 略称データ	DBpedia Japanese	63
DBpedia 地名・施設名 別名データ	DBpedia Japanese	187

表 4: 評価用データの例

地名・施設名	意味 ID
東京大学	1
東大	1
都留 文化 大	2
都留 文 大	2
都留 文	2
…	…

6. 実験結果

既存手法と提案手法の評価結果を図 5 に示す．提案手法である SS と SS+SG では，既存手法 SG と比較し，全ての評価データにおいて MRR が向上していることがわかる．MRR の向上幅は「DBpedia 地名・施設名 別名データ」においても最も高く，既存手法 SG の 0.151 に対し，提案手法 SS+SG では 0.418 と，177%向上している．–なお，各実験結果に対し，ウィルコクソンの符号付順位和検定を適用したところ，「大学の略称データ」「DBpedia 地名・施設名 別名データ」において，既存手法 SG と提案手法 SS, SS+SG の間において有意水準 1%で逆順位の母平均に有意差が認められた．これらの結果から，提案手法では既存手法と比べて，地名・施設名の「意味」の獲得精度が高まっていることがわかる．

入力クエリを「都留 文 大」とした場合の例を表 5 に示す.「都留 文 大」の同義語である「都留 文」（共に正式名称は「都留文科大学」）が，既存手法では，全 259 件中第 98 位に現れるのに対し，提案手法では第 1 位に現れており，既存手法と比べ，提案手法では「都留 文 大」の意味の獲得精度が高いことがわかる. また, 既存手法では, 第 1 位が「大 大 大」であり（正式名称は「大阪大谷大学」),「大」を多く含む地名・施設名であること，検索結果の上位 5 件全てが「大」を含むことがわかる．これは，既存手法では「都留」の意味を学習できておらず，そのため，「大」の意味に強く引っ張られた検索結果になったためと考えられる．これに対し，提案手法では，同義語である「都留 文」が第 1 位に現れているだけでなく，「都留文科大学」の所在である山梨を所在地とする大学である「山梨学院大」が第 2 位に来ている．これは，提案手法では空間情報を考慮した分散表現が獲得できているためと考えられる．

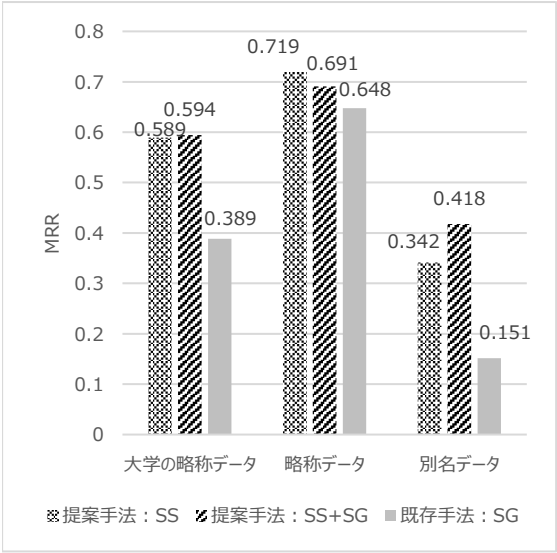

図 5: 既存手法と提案手法の評価結果．提案手法である SS と SS+SG では，既存手法 SG と比較し，全ての評価データにおいて MRR が向上している

表 5: 入力クエリ「都留 文 大」に対する地名・施設名の検索結果．括弧内は正式名称．半角スペースは単語の区切りを表す．「都留 文 大」の同義語である「都留文」（共に正式名称は「都留文科大学」）が，既存手法では，全 259 件中第 98 位に現れるのに対し，提案手法では第 1 位に現れており，既存手法と比べ，提案手法では「都留 文 大」の意味の獲得精度が高いことがわかる．

順位	提案手法：SS+SG	既存手法：SG
1	**都留 文** （都留文科大学）	大 大 大 （大阪大谷大学）
2	山梨 学 大 （山梨学院大学）	学芸 大 （東京学芸大学）
3	旭 教 大 （北海道教育大学旭川校）	弘 大 （弘前大学）
4	旭 教 （北海道教育大学旭川校）	大 歯 大 （大阪歯科大学）
5	東 工科 大 （東京工科大学）	大 薬 大 （大阪薬科大学）

7. まとめ

本研究では，施設名や地名の検索への応用を目的に，空間座標情報を用いた単語の分散表現獲得手法を提案した．既存の分散表現獲得手法では，単語の出現文脈に基づき分散表現を獲得する．しかし，既存手法では，空間の情報を考慮しないため，異なる対象を表す分散表現が類似する問題がある．これに対して提案手法では，自然言語で記述された名称とその空間座標情報からなる大規模データから分散表現を学習する. これにより, 異なる対象については，相違した分散表現が得られる．評価は同義語検索タスクにより行った．その結果，提案手法は，既存手法と比較し，全ての評価データにおいて MRR が向上することがわかった．また，MRR の向上幅は「DBpedia 地名・施設名 別名データ」においても最も高く, 既存手法の 0.151 に対し, 提案手法では 0.418 と，177%向上した.

今後の課題を述べる．本研究では，空間座標情報としてウィキペディア日本語版から取得した地名・施設名とその緯度・経度情報を用いた．今後は，緯度・経度情報だけでなく，三次元の座標情報や，時間を考慮した座標情報を利用することで，工場における設備名や設計図面における部品名の検索などに応用したい.

参考文献

[1] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, In ICLR (2013)

[2] 城光英彰, 松田源立, 山口和紀.: 文脈限定 Skip-gram による同義語獲得, 自然言語処理, Vol. 24, No. 2, pp. 187-204 (2017)

[3] Zellig, H.: Distributional structure, Word, Vol. 10, No. 23, pp. 146-162 (1954)

[4] 城光英彰, 松田源立, 山口和紀.: 同義語判定問題を用いた語義ベクトルの評価の検討, 第 10 回 インタラクティブ情報アクセスと可視化マイニング研究会 (2015)

[5] Caruana, R.: Multitask Learning, Machine Learning, vol.28 (1997)

[6] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” NAACL. 2018.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” CoRR, abs/1810.04805, 2018.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” NIPS. 2017.

[9] Tian, R., Okazaki, N. and Inui, K.: The mechanism of additive composition, CoRR, Vol. abs/1511.08407（2015）

[10] Asahara, M., and Matsumoto, Y.: ipadic version 2.7. 0 User’s Manual, Computational Linguistics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology (2003)

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in Advances in Neural Information Processing Systems (2013)