2I4-OS-17a-5

分散表象とオントロジーの関係

Relationships Between Distributed Represenation and Ontology

市瀬 龍太郎 *1　荒川 直哉 *2
Ryutaro Ichise　Naoya Arakawa

*1国立情報学研究所　*2ドワンゴ人工知能研究所
National Institute of Informatics　Dwango Artificial Intelligence Laboratory

In this paper, we analyzed relationships between distributed representation created by neural network language model and ontology in order to construct a method to create is-a relationship on ontologies. We conducted experiments to evaluate the method. The experimental results show that we can construct is-a relationships partially by using the proposed method.

1. はじめに

汎用人工知能を実現するためのアプローチとして，近年，脳型の計算手法が注目を集めている．脳型の計算手法においては，一般的に，知識がニューロン同士の結合の強さなどによって表現され，様々な場所に分散して保存される．これは，分散表象と呼ばれ，表出される知識との関係が掴みづらいという問題点があることが知られている．人間レベルの人工知能を脳型の計算を用いて実現する場合には，分散表象と表出される知識との間の変換を行う必要がある．なぜならば，汎用人工知能が知識を獲得していく最も有効な方法の一つは，他のエージェントから獲得することであり [Langley 09]，そのためには，自分の内部で持つ分散表象の知識を他のエージェントが分かるような形態で，表出しなければならないからである．その問題を解決することは，脳型の計算手法を実用化していく際に，大きな鍵となるであろう．そこで，本論文では，分散表象と明示化された知識の一形態となるオントロジーの間をつなぐ手法について考察を行う．

分散表象と明示化された知識をつなぐ方法の一つとして，ニューラルネットワーク言語モデルがある．ニューラルネットワーク言語モデルでは，大量の文章を用いて，入力単語と出力単語の関係を学習し，単語を高次元ベクトルで表現して保持する．本研究では，そのモデルを用いることで，高次元ベクトルで表現された単語の知識から人間が用いる概念体系に近いオントロジーを導出する方法について検討を行う．

2. ニューラルネットワーク言語モデル

ニューラルネットワーク言語モデルは，ニューラルネットワークを用いて，単語の高次元ベクトルと統計的言語モデルの学習を行う．ニューラルネットワーク言語モデルとして，Bengio らのモデル [Bengio 03] や，Mikolov らのモデル [Mikolov 13] が知られている．Bengio らのモデルでは，入力単語として，出力単語 W_t の直前に出現する語 $W_{t-1}, ..., W_{t-n+1}$ を利用して，ニューラルネットワークの学習を行い，出力単語の予測を行う．Mikolov らの skip-gram モデルでは，入力単語 W_t を用いて，出力単語 $W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2}$ の予測を行う．skip-gram モデルは，Bengio らのモデルに比べると意味的な精度の高い表現を抽出可能である [Mikolov 13]．そのため，本稿では，単語の高次元ベクトル表現を抽出するために，skip-gram モデルを用いる．

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，E-mail:ichise@nii.ac.jp

3. 単語のベクトル表現とその関係

skip-gram モデルで得られる単語のベクトル同士は，単語間の関係を保持することができる．例えば，Mikolov らは，フランス-パリという関係から，イタリア-ローマ，日本- 東京といった関係を抽出することができると述べている [Mikolov 13]．本論文では，このような関係を構造化することで，オントロジーのように明確な知識構造を構築することを試みる．

ある単語 W_i の単語ベクトルを V_{W_i} とする．2 つの単語 W_1, W_2 が与えられた時に，2 つの単語の間の特定の関係 $R(W_1, W_2)$ は，以下の式で表される．

$$R(W_1, W_2) = V_{w_1} - V_{w_2}$$

その時に，単語 W_1, W_2 がそれぞれある概念を表すとすると，$R(W_1, W_2)$ は，概念間の関係を表現するベクトルとなる．これを利用すると，ある新たな概念を表す単語 W_3 が与えられた時に，W_3 に対して，W_1, W_2 と同じ関係にある単語 W_4 のベクトル V_{W_4} は，以下の式により表現することが可能となる．

$$V_{W_4} = V_{W_3} - R(W_1, W_2)$$

オントロジーの定義する際に，様々な概念間の関係が使われるが，本論文では，オントロジー定義の際に，基本となる is-a 関係に注目する．is-a 関係は，概念の上下関係を表す．そのため，is-a 関係をベクトル空間上で表す関係 R を発見できれば，ある概念を表す単語 W_x の上下の概念を見つけることが可能となる．

これまでに，様々なオントロジーが構築されてきた．それらを用いることにより，既知の is-a 関係を利用することができる．本研究では，既知の is-a 関係を利用することで，未知の is-a 関係を発見し，それを利用することでオントロジーを構築することを提案する．なお，ここで用いる既知のオントロジーは，全てのクラスが定義されている必要はなく，is-a 関係が網羅されている必要もない．

データ	100 位以内の数	順位の合計	平均の順位
データ 0	173	4384	25.34
データ 1	159	4362	27.43
データ 2	134	4225	31.53
データ 3	158	3981	25.20
データ 4	159	4283	26.94
データ 5	162	4631	28.59
データ 6	146	4071	27.88
データ 7	160	4029	25.18
データ 8	162	4141	25.56
データ 9	168	3683	21.92
平均	158.1	4179.0	26.43

表 1: 実験結果.

4. 実験

4.1 実験設定

本研究で提案した手法を用いて，オントロジーを構築するために，is-a 関係がどの程度の精度で抽出できるか評価する実験を行った．実験では，まず，skip-gram モデルを用いて，単語ベクトルの生成を行った．単語ベクトルを生成する際には，単語ベクトルの次元数を 300 次元に設定し，コーパスとして英語版のウィキペディア，実装として gensim [*1] を用いた．

オントロジーにおける is-a 関係の評価のために，WordNet を利用した．WordNet では，is-a 関係が記述されているが，多義語が含まれている場合がある．そのため，WordNet の全ての Synset の中から，数字や記号が含まれている単語を取り除き，意味が 1 つのみを含む単語を抽出した．さらに，その中で，is-a 関係にあり，skip-gram モデルで単語ベクトルを抽出できた単語のみを対象とした所，6800 組の is-a 関係を取り出すことができた．

この 6800 組のデータに対して，10-fold の交差検定法を用いて，未知の is-a 関係が抽出できるかを調べた．まず，6120 組の単語を用いてそれぞれに，$R(W_1, W_2)$ を計算し，その平均を is-a 関係を表す関係ベクトル R とみなした．そして，その関係ベクトルを利用して，残りの 680 組の単語の子の単語 W_3 から，親の単語 W_4 の予測を行った．そのために，W_4 に相当するベクトル V_{W_4} の計算を行った後に，近傍にある単語を is-a 関係にある親の単語とみなした．これを全ての fold に対して繰り返し，全ての単語の親に相当する単語を取り出した．

4.2 実験結果

親単語のベクトルの最近傍の単語を抽出した場合に，is-a 関係を適切に取り出すことができたものは皆無であった．そのため，最近傍という制約を弱め，上位 100 位以内に入っている単語を改めて調べた．その結果を表 1 に示す．表では，データ中で，親の単語が上位 100 位以内に出現した単語の組の数，100 位以内に出現した場合の順位の合計，100 位以内に出現した場合に，何番目に出現したかの平均値を掲載した．

10-fold の交差検定を用いたため，10 個のデータがある．それぞれのデータは，680 個のテストデータを含む．表 1 より，100 位以内に適切な単語が入っている平均は，158.1 個となっているため，4 分の 1 弱のデータに対して，上位 100 個以内で親の単語を抽出できていることが分かる．一方，順位に着目すると，平均で 26.43 位となっている．もし，ランダムで出現すると仮定すると，平均順位は 50 位になることが期待されるため，本手法により，より適切に，is-a 関係にある単語を抽出できることが分かる．

∗1 https://radimrehurek.com/gensim/

データを詳細に見ていくと，興味深いことが発見された．まず，親になる単語ベクトルの最近傍の単語を見ると，ほとんどが自分の単語を示していた．つまり，is-a 関係を表す関係 R が，小さなベクトルとして設定されていることが分かる．今回の実験では，関係 R を計算する際に，6120 個の is-a 関係を用いて，その平均値を用いた．WordNet では，is-a 関係として記述されるものに，誤ったものがあることが知られている [Guarino 98]．例えば，オントロジーの設計時には，タイプとロールを分けて，is-a 関係を作る必要があるが，WordNet ではそれらを混ぜて設計している．そのような複数の関係を利用して，関係 R を計算しているため，結果として，関係 R に明確な方向性が出なかった可能性がある．そのような場合の対策として，関係 R をいくつかの類似した関係に分けて抽出し，利用するということが考えられる．

また，今回の実験では，得られた結果が学習データにより，バイアスがかかった可能性があることも示唆された．ウィキペディアの記事は，クラスに関する情報よりもインスタンスに関する情報の方が詳細に記載されている．そのため，概念を学習する際に，適切な語で十分に記述されていないことが考えられる．例えば，gook(東洋人) という単語の場合には，WordNet で親の単語は oriental であるが，この単語は，他の部分も含めてほとんど出現していない．代わりに，多くの人名が gook に対して，候補として上がっていた．また，peba(ココノオビアルマジロ) という語は，armadillo が親の概念となるが，ウィキペディアでは，アルマジロとして全く記述されていない．そのため，適切な概念を学習するための学習データをどうするかについても，深く考えていく必要があるであろう．

5. まとめ

本研究では，ニューラルネットワーク言語モデルを用いた分散表象と代表的な知識表現であるオントロジーの関係について考察を行った．その考察に基づき，オントロジーの表現に必要となる is-a 関係を分散表現からある程度構築できることが実験により示された．今後は，is-a 関係の分類を行い，詳細な実験をすることや，オントロジー構築目的に沿った学習データの選び方などが課題になると考えられる．

参考文献

[Bengio 03] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C.: A Neural Probabilistic Language Model, *Journal of Machine Learning Research*, Vol. 3, pp. 1137–1155 (2003)

[Guarino 98] Guarino, N.: Some Ontological Principles for Designing Upper Level Lexical Resources, in *Proceedings of the 1st International Conference on Language Resources and Evaluation* (1998)

[Langley 09] Langley, P., Laird, J. E., and Rogers, S.: Cognitive architectures: Research issues and challenges, *Cognitive Systems Research*, Vol. 10, No. 2, pp. 141–160 (2009)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, in *Proceedings of Workshop at International Conference on Learning Representations* (2013)