

ベクトル空間表現を用いた JNL2KR の語義曖昧性解消

檜原和昭¹ 植松すみれ^{2,4} 宮尾祐介^{2,3,4} Chitta Baral¹¹ アリゾナ州立大学 ² 国立情報学研究所 ³ 総合研究大学院大学 ⁴ JST さきがけ

{kkashiha, chitta}@asu.edu, {uematsu, yusuke}@nii.ac.jp

1 はじめに

人工知能の分野には、様々な目標を達成するための種々多様な知識表現 (KR) 言語がある。例えば、一階述語論理は数学的な知識を表現することに用いられ、時相論理はエージェントやロボットの目標の表現に有用であり、そしてデータベース知識にアクセスするためのデータベースクエリ言語がある。これらの KR 言語は、自然言語からこれらの望ましい KR 言語のいずれかに最もよく翻訳するという課題を提起している。

意味的構文解析器は自然言語から KR 言語の表現へ翻訳する方法の 1 つであり、自然言語からあらゆる形式の KR 言語へ翻訳するための一般的な方法を確立した様々な意味的構文解析システム [15, 5, 7, 3] が提案されている。幾つかのシステムは手作業で作成したルール [16, 6] を利用して語彙の意味を学習する一方、NL2KR [1, 4, 12] や JNL2KR [17] のようなシステムは逆ラムダ計算を用いて語彙の意味を学習する。これらのシステムの共通する課題の 1 つが、未知の単語の扱いである。自然言語から KR 言語への一般的な翻訳の方法を確立しようとする意味的構文解析器にとって、未知の単語の意味を学習する確かな方法を持つことは重要である。最も一般的な手法は Generalization [1] として知られており、既知の意味表現に基づいて未知の単語に意味のテンプレートを適用することで未知の単語の意味を学習する。Generalization の問題としては、テンプレートを適用する際に未知の単語とそのテンプレートの元となった既知の単語の類似度を考慮しないため、不要な意味を学習することがある。

今回、我々は意味的構文解析で未知の単語の取り扱いを改良することを考える。ベクトル空間表現と新しい 3 つ組の意味テンプレートセットを元に新たな Generalization モジュールを実装することで、未知の単語と既知の単語間の語義曖昧性解消し、最も似た既知

の単語から意味を学習する。この新しい Generalization モジュールを、日本語文から任意の KR 言語へ翻訳する辞書をラムダ計算を用いて学習する意味的構文解析プラットフォームである JNL2KR に適用し、従来より小規模かつ精度の高い辞書を学習することで精度向上を図る。

ベクトル空間表現として Word2Vec (W2V) [8] を使用し、標準的なデータセット GeoQuery250 [13] を用いて評価実験を行った。評価実験の結果は、JNL2KR+W2V が従来の約 7 割のサイズの出力辞書を学習し、その精度の高い辞書から最先端の精度を得られることを示している。

2 研究背景

JNL2KR [17] は、日本語文から意味表現へ翻訳するシステムを半自動で構築する初のプラットフォームである。このプラットフォームは、単語と句の持つ意味をラムダ計算表現として表現する Montague の手法 [9] を用いており、ラムダ計算表現では文の意味は適切なラムダ計算 [2] の関数適用を介して構成単語の意味から構築する。JNL2KR は学習モジュールと翻訳モジュールに分かれている。

学習モジュールでは初期辞書と翻訳付きの例文を元に、CCG パーザで例文を解析し、未知の単語の意味を逆ラムダ計算及び Generalization [1] で学習することで、初期辞書よりも多くの単語の意味を収録した辞書を出力する。例えば、“太郎はご飯を食べる。eat(太郎, ご飯)”という例文と“太郎 NP 太郎”(単語: 太郎, CCG カテゴリ: NP, 意味表現: 太郎), “。S\ S $\lambda x.x$ ”, “は NP\ NP $\lambda x.x$ ” 及び “を NP\ NP $\lambda x.x$ ” という初期辞書が与えられた時、図 1 に示すように、Generalization で“ご飯”の意味が“太郎 NP 太郎”の例から“ご飯 NP ご飯”と学習され、逆ラムダ計算で“ご飯を食べる”及

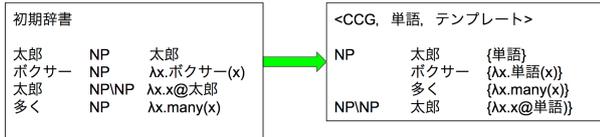


図 3: 提案手法で用いる 3 つ組のテンプレートセットの例.

似度を用いる. 今回, 単語のベクトル空間表現として Word2Vec を用いるが, フレームワーク自体は柔軟であり, 他のベクトル空間表現にも対応している.

前述の花子の例の場合, ベクトル空間表現として Word2Vec を用いると, 花子と同じ CCG カテゴリの単語は“太郎”, “ボクサー”と“最も多く”である. ここで, “太郎”と“花子”の類似度は 0.56, “ボクサー”と“花子”の類似度は 0.25, そして“多く”と“花子”の類似度は 0.22 であるから, “太郎”が最も類似の単語となる. 図 3 より, “太郎 NP”のテンプレートから“花子 NP 花子”が学習される. この例が示すように, 従来の Generalization のように有用でない意味を学習することが減ると期待される.

4 評価実験

JNL2KR との直接比較を行うため, 日本語訳の GeoQuery250 コーパス [13] を用いて我々の手法を評価する. GeoQuery250 コーパスは 250 のアメリカ合衆国の地理に関する日本語音訳の質問文とその意味表現をペアとして持っている. また, この意味表現は質問の解をアメリカ合衆国の地理情報データベースから抽出できる表現となっている. 我々は, JNL2KR で行われた評価実験と同様に, 250 文の元の英文から日本語へ日本語翻訳家が翻訳したデータを用いる.

その他の関連研究 [13, 14, 6, 7] との直接比較も行う為, 固有名詞はカテゴリ毎に単一の単語 (例: 州名は全て STATE, 都市名は全て CITY) に変換し, 関連研究の実験で用いられた標準的な 10 回交差検定を用いて評価する. 適合率, 再現率, F1-尺度を報告する. 適合率は出力された論理形式が正しい割合, 再現率は正しい論理形式が出力された文章の割合, F1-尺度は適合率と再現率の調和平均である. 我々は, テスト文の実際の論理形式と我々のシステムの出力が完全一致した場合, 翻訳は正しいと判断する.

本実験は初期辞書が必要であるが, JNL2KR の実験で使用された初期辞書を我々の評価実験でも使用する.

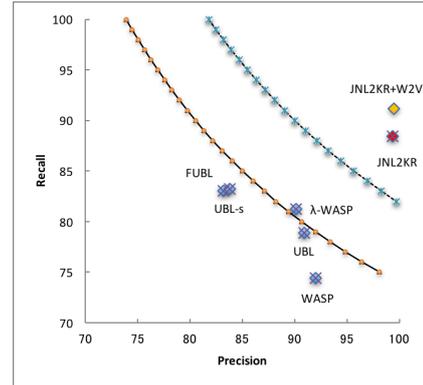


図 4: GeoQuery250 の適合率 (Precision), 再現率 (Recall), そして, F1-尺度の比較. 破線は F1 値が 90% を示し, 実線は F1 値が 85% を示す.

システム (%)	再現率	適合率	F1 値
WASP	74.4	92.0	82.9
λ -WASP	81.2	90.1	85.8
UBL	78.9	90.9	84.4
UBL-s	83.0	83.2	83.1
FUBL	83.2	83.8	83.5
JNL2KR	88.4	99.4	93.5
JNL2KR + W2V	91.1	99.5	95.1

表 1: 日本語 GeoQuery250 データセットの完全一致率
初期辞書の大きさは 250 語である. 250 語のうち, 31 語が接尾辞, 5 語が記号または接続詞, 75 語が動詞か助動詞, 47 語が形容詞か形容動詞, そして 92 語が名詞または名詞句である. この中には同じ単語で複数の異なる意味を持つ場合も含まれる.

図 4 及び表 1 は関連研究と GeoQuery250 データセットでの JNL2KR 及び我々の JNL2KR+W2V の精度比較を示している. 表 1 は, 変数なしの意味表現を用いた WASP の精度, ラムダ計算の意味表現を用いた λ -WASP, UBL, UBL-s 及び FUBL の精度を示す.

表 1 が示すように, JNL2KR+W2V が全てにおいて最も高い精度を示した. また, 学習モジュールの出力辞書のサイズは約 30% の減少が認められ, 新たに学習した単語の数が約 60% の減少した. 例えば, “最大”は“の”と組み合わせられて形容詞として用いられるが, CCG カテゴリが NP である. JNL2KR では必要な形容詞の意味 ($\lambda z. \lambda y. \lambda x. \text{max}(x, z@x, y@x)$ など) のみならず, 初期辞書に与えられている NP の名詞 (STATE NP stateid(STATE)) などから名詞の意味 (最大 NP 最大 id(STATE) など) が混在して学習されていた. しかし, 我々の手法では形容詞として用いられる意味のみ

学習していた。CCG カテゴリが *NP* の単語が名詞のみならず形容詞なども含まれているため、JNL2KR の *Generalization* では同じ CCG カテゴリ内の全ての品詞の意味が学習されるなどの問題があったが、改良した *Generalization* は最も類似した既知の単語の意味テンプレートから未知の単語の意味を学習するため、類似語が同じ品詞となる場合が多く、適切なテンプレートのみを用いて学習できたと考えられる。以上の結果から、JNL2KR+W2V は従来よりも小規模かつ精度の高い出力辞書を学習することができることを示した。

5 おわりに

日本語文から同等の意味表現へ変換する翻訳システムを構築する為のプラットフォーム JNL2KR の *Generalization* を単語のベクトル空間表現を用いて改良した。我々の提案手法は一般的な *Generalization* の拡張であり、英語版の NL2KR へも応用が可能である。

ベクトル空間表現として Word2Vec を用いた評価実験を行い、出力される辞書のサイズの減少と精度の向上が認められた。本提案手法が未知の単語に対して最も類似の既知の単語から有用な意味のみ学習することで、従来の手法より精度の高い辞書を得ることができたためと考えられる。

今後の課題としては、他のデータセットによる評価実験および GloVe [10] や他のベクトル空間表現との比較などが挙げられる。

参考文献

- [1] C Baral, J Dzifcak, K Kumbhare, and N H Vo. The NL2KR system. In *Proceedings of LPNMR 2013*, 2013.
- [2] A Church. An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, 1936.
- [3] L Dong and M Lapata. Language to logical form with neural attention. *arXiv preprint arXiv:1601.01280*, 2016.
- [4] S Gaur, N H Vo, K Kashihara, and C Baral. Translating simple legal text to formal representations. In *JURISIN 2014*, 2014.
- [5] R Ge and R J Mooney. Learning a compositional semantic parser using an existing syntactic parser. In *Proceedings of the ACL-IJCNLP 2009*, pp. 611–619. ACL, 2009.
- [6] T Kwiatkowski, L Zettlemoyer, S Goldwater, and M Steedman. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of EMNLP 2010*. ACL, 2010.
- [7] T Kwiatkowski, L Zettlemoyer, S Goldwater, and M Steedman. Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of EMNLP 2011*, 2011.
- [8] T Mikolov, K Chen, G Corrado, and J Dean. word2vec, 2014.
- [9] R Montague. English as a Formal Language. In *Formal Philosophy: Selected Papers of Richard Montague*, pp. 188–222. Yale University Press, 1974.
- [10] J Pennington, R Socher, and C D. Manning. Glove: Global vectors for word representation. In *EMNLP 2014*, 2014.
- [11] M Steedman. *The Syntactic Process*. MIT Press, Cambridge, MA, 2000.
- [12] N H Vo, A Mitra, and C Baral. The nl2kr platform for building natural language translation systems. In *Proceedings of ACL*, 2015.
- [13] Y W Wong and R J. Mooney. Learning for semantic parsing with statistical machine translation. In *Proceedings of HLT-NAACL 2006*, 2006.
- [14] Y W Wong and R J. Mooney. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of ACL 2007*, 2007.
- [15] L S. Zettlemoyer and M Collins. Online learning of relaxed ccg grammars for parsing to logical form. In *In Proceedings of EMNLP-CoNLL-2007*, 2007.
- [16] L S Zettlemoyer and M Collins. Learning context-dependent mappings from sentences to logical form. In *Proceedings of ACL and AFNLP*, 2009.
- [17] 檜原和昭, 植松すみれ, 宮尾祐介, Chitta Baral. JNL2KR システム 用いた日本語から意味表現への変換. 言語処理学会 第 22 回年次大会, 2016.