修士学位論文要約（平成29年３月）

分散表現による外部知識の自然言語解析への適用

小松　広弥
指導教員：乾　健太郎

Applying External Knowledge in Distributed Representations for Parsing Natural Language

Hiroya KOMATSU
Supervisor: Kentaro INUI

Given sentences in natural language, parsing obtains their syntactic and semantic structures. It is the foundation of many NLP tasks. Generally, a parser is constructed by a supervised learning. In order to improve its performance, it is necessary to design features and to increase the amount of the training data. In parsing, lexical features are very powerful and commonly used, but they tend to be overfitting especially when they scarcely appear in the training data. In addition, it is hard to manually annotate a large amount of training data because of the complexity of the structures. In this study, we propose embedding features in parsing to remedy the disadvantage of lexical features. In semantic parsing, we propose to add partial training data based on the framework of the semantic representations. In experiments, we show improvements of the performances of dependency parsing and semantic parsing. Furthermore, our analyses show the advantages of the embedding features and some interesting problems of semantic parsing.

1. はじめに

自然言語解析とは，自然言語の文から文が持つ構文的構造や意味的な構造を得るものであり．様々な自然言語処理タスクの入力として利用するためその精度が重要である．自然言語解析は教師あり学習によって構築されるが，その素性には単語文字列の異なりに注目した単語表層素性がよく用いられる．この素性は非常に強力であるが訓練時に未知の単語に対して学習できないことや既知の単語に対して過学習してしまうことが問題として挙げられる．また意味表現解析においては，意味表現の複雑さから多量の訓練データを人手で用意することはコストが高いことが問題として考えられる．

本論文では，まず単語分散表現による素性を提案する．単語分散表現は類似単語同士で近いベクトルを持つ性質があるため，解析器に類似単語の情報を与えられる．また，意味表現解析において，意味表現が土台としているフレームワークを反映したアノテーション付テキストを新たな訓練データとして加えることを提案する．

2. 分散表現素性の依存構文解析への利用

依存構文解析は文内の単語間の修飾関係を求める解析である．依存構文解析では単語表層素性が非常に効果的である[1]が，1 つの単語表層素性はベクトルで表示したとき，ある要素だけが 1 でその他の要素が 0 の one-hot なベクトルとなっている．そのため，単語表層素性に対する重みは単語ごとに異なる値が発火し，また訓練データにない単語に対しては重みベクトルの値が学習されない．

・単語表層素性 (語彙数次元)

saw　　= (0, 0, …, 0, 1, 0,…, 0)

watch　= (0, 0, …, 0, 0, 1,…, 0)

・分散表現素性 (d 次元)

saw　　= (0.21, 0.56)

watch　= (0.22, 0.57)

図 1：分散表現素性の概念図

本手法では単語表層素性の one-hot ベクトルを単語分散表現で置換することを提案する．One-hot ベクトルを対応する単語分散表現で置き換えたものを分散表現素性とする(図 1)．単語分散表現を素性に用いることにより類似する単語同士で解析器のスコアが類似し，また未知語に対してもスコアが計算される．組合せ素性に関しては，2 ベクトルの外積，加算，要素積，結合を比較する．また単語分散表現を構築する際の文脈単語の位置についても比較を行う．

実験では Huang ら[1]の Shift-Reduce 型解析器を再実装し，全ての単語表層素性を分散表現素性に置換した．Penn Treebank を学習データとしてベースラインと提案する最良の手法の性能を表 1 に示す．提案手法はベースラインに対して性能向上が

確認でき，また未知語を含む文に対して顕著な性能向上があった．分散表現素性の利点として，実際に未知の単語を含む文に対する効果と類似単語同士で解析器の動作を類似させる効果があることを分析した．

正答率(%)	開発	テスト	未知
baseline	91.93	91.67	89.01
提案モデル	92.57	92.20	90.27

表 1：依存構文解析の性能

3. オントロジ情報の意味解析への利用

Abstract Meaning Representation (AMR)は，「誰が何に対して何をしたか」を表現できる意味表現である．AMR は Proposition Bank (PropBank)と呼ばれるオントロジに基づいている．PropBank は述語の種類とその項の定義と，文に述語項の関係を付与したものを提供している．AMR は PropBank の定義を利用して，述語項関係や修飾関係，固有表現などあらゆる意味関係を表している．

本研究では，AMR 解析の訓練データとして，PropBank のアノテーション付テキストを利用することを提案する．アノテーション付テキストは述語項関係のみを表示した部分的 AMR 付訓練データとして捉えることができ，AMR の訓練データに加えて訓練エータとは他のアノテーション付テキストを学習に加えることによって，述語項関係に対しての学習量を増加させ，ロングテイルな学習が可能になると考えられる．また，2節で提案した分散表現素性も同時に利用することで述語項関係の汎化を狙う．

PropBank アノテーション付テキストを学習に加えるためには，解析器[2]の動作に基づいてアノテーション付テキストを訓練データに変換する．AMR 解析器は述語の種類を特定する判定器や項を特定する判定器などに分離することができるので，それぞれの判定器の学習できる部分だけをアノテーション付テキストから学習を行う．

実験では，Wang ら[2]が公開している実装を利用し，PropBank アノテーション付テキストを学習に加えたときと，分散表現素性(embed)も同時に利用したときで性能の比較を行った．評価値は先行研究[3]に従う．その結果を表 2 に示す．テストセットに対する性能向上から，訓練データとは他の文に対してより高性能に解析できていることが確認できた．

評価値[3]	開発	テスト
baseline	65.67	63.15
baseline + PropBank	65.66	64.00
baseline + Propbank + embed	65.66	64.56

表 2：AMR 解析の性能

分析では，述語，関係別の解析性能から PropBank アノテーション付テキストは述語や項の種類の判定に対して効果が大きいこと，また判定器毎の性能から，分散表現素性は AMR に特有の構造に対しての解析に効果が大きいことを確認した．

解析を誤った例として，”*With this bridge, the distance would be very small.*”という文を考える．この文の AMR には cause-01(ARG0: bridge, ARG1: small)「bridge が small を引き起こす」という述語項関係が表現されている．この文から”cause-01”を解析するためには，”will be A with B ”が”B cause A” という意味を持つといったような，単語の組合せによる意味の変化を捉える必要がある．しかし既存の素性や PropBank，分散表現素性からこの意味の変化を得るのは難しい．今後の課題としては，このような構成的な意味の変化を捉えられる素性を設計することが考えられる．

4. まとめ

本論文では，頑健で汎用的な自然言語解析器の構築のために，まず単語分散表現を用いた分散表現素性を提案した．この分散表現素性を依存構文解析に対して適用したところ，依存構文解析の性能向上に寄与することを確認した．この分散表現素性を用いる利点として，単語分散表現が (1) 訓練時に未知の単語を既知の単語と結びつける効果，(2) 類似する単語同士で解析器の動作を類似させる効果があることを分析した．

次に，意味表現解析である AMR 解析において，AMR の背後にあるオントロジの情報を部分的な AMR 構造として学習データに取り入れることで，AMR 解析の性能向上に寄与することを確認した．AMR 解析においても本論文で提案した分散表現素性を用いたとき，さらなる性能向上が確認できた．AMR 解析における PropBank による学習は概念や関係のラベル付などに，また分散表現素性は，AMR 解析に特有な処理である単語を除去したり結合したりする処理などに貢献することを分析した．さらに，AMR 解析が誤った事例を定性的に分析し，意味的な構造を捉える場合は単語の組合せによる意味の変化を捉える必要があり，AMR 解析の今後の課題を考察した．

[1] Liang Huang, Suphan Fayong, and Yang Guo. Structured perceptron with inexact search. In Proceedings of NAACL-HLT, 2012

[2] Chuan Wang, Sameer Pradhan, Xiaoman Pan, Heng Ji, and Nianwen Xue. CAMR at semeval-2016 task 8: An extended transition-based AMR parser. In Proceedings of SemEval-2016, pp. 1173–1178, 2016.

[3] Shu Cai and Kevin Knight. Smatch: an evaluation metric for semantic feature structures. In Proceedings of ACL, pp. 748–752, 2013.