

(続紙 1)

京都大学	博士（情報学）	氏名	王 荀（Wang Xun）
論文題目	Entity-Centric Discourse Analysis and Its Applications （エンティティに注目した談話解析とその応用）		
（論文内容の要旨）			
<p>This dissertation presents an entity-centric view of discourse analysis and investigates its applications in several important tasks including automatic summarization, reading comprehension and machine translation. The entity-centric view regards entities as the core of discourse and all other parts of discourse are elaborated around entities. The theoretical background and the method of entity-centric discourse analysis are explained in this dissertation. The contents of this dissertation are organized into 8 chapters.</p> <p>Chapter 1 introduces discourse analysis and provides an overview of the proposed entity-centric method. An outline of this dissertation is also included.</p> <p>Chapter 2 discusses the history and development of discourse analysis, in particular, the theoretical development of discourse analysis. Concise introductions to two important theories – structuralism and generative grammar – are presented. The two theories have had huge impacts on discourse analysis, and they serve as the theoretical guidelines of the entity-centric method. The motivation for using the proposed entity-centric method is also explained.</p> <p>Chapter 3 and Chapter 4 discuss how to learn word representations and how to discover hidden entities in discourse. They serve as the basis of the entity-centric representation for discourse.</p> <p>Chapter 3 presents the work on learning word embedding. As pointed out in structural linguistics, discourse is a system composed by elements that are defined by their mutual interactions. This work organizes discourse into a tree, and nodes in the tree are regarded as being decided by their siblings and their parents. Words of discourse are connected via their common ancestors in the tree. This method thus remedies the problem of existing work, which focuses only on neighboring words and ignores relationships between distant words.</p> <p>Chapter 4 describes the work on discovering hidden nominal words. According to generative grammar, discourse, as is observed, distorts from its original form – the deep structure. The deep structure contains underlying information that is not explicit to computers but vital for discourse analysis. The hidden nouns are part of the unobserved information. A neural network approach, which uses tree-based features, is proposed to detect those hidden nouns in discourse. Experiments on Chinese Treebank show the effectiveness of the proposed model. When evaluated by F-1 scores, a significant improvement of more than 10 points is achieved, as compared to previous work.</p> <p>Chapter 5, 6 and 7 discuss applications of the entity-centric discourse analysis.</p> <p>In Chapter 5, an entity-graph-based summarization system is demonstrated. It is assumed that entities not only carry information but also connect sentences to form coherent passages. A good extractive summary can be constructed by selecting a sequence of sentences which contains continuous appearances of important entities. An objective function based on entities is developed</p>			

accordingly. Experiments demonstrate that the extracted summaries are both informative and coherent.

In Chapter 6, an entity-based memory network model is developed for reading comprehension tasks. Information conveyed by discourse is represented using its containing entities. All the entities in a piece of discourse construct a memory pool. The incoming question is answered according to the most related entities in the memory. It is also shown that the proposed model can be easily applied to tasks other than reading comprehension. An increase of about 2 points is obtained compared with previous work when tested on a classification task.

In Chapter 7, the entity-centric analysis of discourse is used for machine translation. A sentence is decomposed into entities and relations, which are translated separately. The meanings of entities are concrete and stable. But relations are abstract and volatile. Two translation models are trained separately, one for entities and the other for relations. The translated entities and relations are then used together to generate the target sentence.

In Chapter 8, a conclusion of the entity-centric discourse analysis method is presented along with discussions relating to future work.

注) 論文内容の要旨と論文審査の結果の要旨は1頁を38字×36行で作成し，合わせて，3,000字を標準とすること．

論文内容の要旨を英語で記入する場合は，400～1,100 wordsで作成し

審査結果の要旨は日本語500～2,000字程度で作成すること．

(論文審査の結果の要旨)

本論文は、談話解析においてエンティティ（名詞で表される事物）を談話の中核とみなし、談話解析の基盤となる単語の意味の分散表現と省略名詞句の復元について、さらにその応用として自動要約、質問応答、機械翻訳について、エンティティ中心の新たな解析手法を提案したもので、得られた主要な成果は以下の通りである。

1. 構造言語学で指摘されているように、談話は階層的な構成要素間の相互作用によって定義される。単語の埋め込み意味表現の学習において、談話を単語、文、段落、文書からなる木構造としてとらえ、木のノードの意味表現を兄弟ノードと親ノードとの関係で定まるものと定義した。このモデル化により、近傍単語のみを考慮する既存研究の意味学習に比べて、長距離の単語間の関係を考慮することを可能とし、類似単語推定等の標準データにおいて既存研究を上回る精度を実現した。
2. 談話においては深層構造から表層表現が生み出される過程で名詞句等の省略が起こるが、そのような暗黙の省略名詞句は談話解析において不可欠な情報である。省略名詞句の検出において、構文の特徴を用いて省略名詞句の存在とタイプを推定するニューラルネットワークモデルを提案した。中国語のTreebankを用いた実験において提案モデルの有効性を示し、既存研究と比較してF1スコアで10ポイント以上の有意な改善を達成した。
3. エンティティ中心の談話解析の応用として、自動要約、質問応答、機械翻訳においてこのアプローチの有効性を示した。自動要約では、エンティティをノードとするグラフ表現を用いて、重要なエンティティの連続的な出現を含む文の並びを選択する方法を提案し、これによって良質な抽出型要約が得られることを示した。質問応答においては、談話の情報をエンティティを単位として表現するメモリネットワークを構成し、メモリ内の最も関連性の高いエンティティによって質問に答えるモデルを提案した。機械翻訳においては、エンティティの翻訳が具体的で安定していることに着目し、文をエンティティと関係に分解し、それらの翻訳を別々に訓練する翻訳モデルを提案し、その有効性を示した。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成29年10月19日実施した論文内容とそれに関連した試問の結果、合格と認めた。

注) 論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成 年 月 日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日： 年 月 日以降