氏　　名	NGUYEN Thi Quy
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1972 号
学位授与の日付	平成29年9月28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Building a Treebank for Vietnamese Syntactic Parsing
論 文 審 査 委 員	主　査　准教授　宮尾　祐介 教授　佐藤　健 准教授　山岸　順一 教授　相澤　彰子　国立情報学研究所 准教授　浅原　正幸　国立国語研究所

(別紙様式 2)
(Separate Form 2)

論文の要旨
Summary (Abstract) of doctoral thesis contents

Treebanks, corpora annotated with syntactic structures, are important resources for researchers in natural language processing, linguistic theory, as well as speech processing. They provide training and testing materials so that different algorithms can be compared. However, it is not a trivial task to construct high-quality treebanks. We have not yet had a proper treebank for such a low-resource language as Vietnamese, which has probably lowered the performance of Vietnamese language processing. In order to alleviate such a situation, this thesis has tackled with two objectives, viz., (1) developing a consistent and accurate Vietnamese treebank and (2) applying our treebank to parsing, a crucial problem in improving the quality of speech processing and natural language processing applications. This study is not only beneficial for the development of computational processing technologies for Vietnamese, a language spoken by over 90 million people, but also for similar languages such as Thai, Laos, and so on.

For the first objective, we propose an annotation scheme for the Vietnamese treebank. In comparison with the previous one (VLSP treebank's scheme), our scheme is better because it can cover and distinguish among various constructions and linguistic phenomena in Vietnamese. We also develop three sets of guidelines corresponding to three annotation layers of our treebank, including: word segmentation guidelines (44 pages), part-of-speech (POS) tagging guidelines (73 pages), and bracketing guidelines (182 pages). Our guidelines contain rules to address the challenges of Vietnamese language. Specifically, we hand-crafted 9 rules for segmenting ambiguous expressions, 34 rules for tagging ambiguous words, and 39 rules for bracketing ambiguous expressions. These guidelines, which are used to train the annotators, are valuable resources that serve the use of the treebank.

In addition to developing the annotation guidelines, we describe other issues of ensuring the annotation quality including an appropriate annotation process, a well-designed process of training annotators, and software tools to support the annotation as well as to control the quality. Inter-annotator agreement, intra-annotator agreement, and accuracy of the developed treebank are higher than 90%, which shows that the annotated treebank is reliable and satisfactory. In comparison with the VLSP treebank, our annotation scheme is more fine-grained than the one of the VLSP treebank. For example, our POS tag set includes 33 tags, while there are 17 tags in the VLSP treebank. However, our treebank gives the higher performance in comparison with the VLSP treebank on all of the tasks, namely, automatic word segmentation, POS tagging, and parsing. This indicates that our treebank is more consistent than the VLSP treebank.

For the second objective, we first evaluate representative parsing models on the Vietnamese treebank. We then investigate the errors produced by the parsers and find the reasons for them. Our analysis focuses on four possible sources of the parsing errors, viz., limited training data, word segmentation errors, confusing POS tags, and ambiguous constructions. We use an analysis method that combines the advantages of automatic tools and a manual analysis. While automatic analysis can be applied to a large amount of parsing output, manual investigation can capture the

reasons of the parsing errors precisely. As a result, we find that parsing models based on conditional random field (CRF) and neural network are good for Vietnamese. On the other hand, the quality of Vietnamese parsing can also be improved through enriching contextual information, such as using the hierarchical state-splitting for unlexicalized parsing or exploiting the rich input features of the surface spans for CRF parsing. However, these performances (about 72% in F-score) of Vietnamese parsing are still far lower than the performances reported for English (about 90% in F-score) and Chinese (about 86% in F-score). This indicates that existing models cannot capture contextual information like words working as prefixes and suffixes.

The investigation of parsing errors has revealed the frequent errors in Vietnamese parsing that are VP attachment, NP attachment, PP attachment, and clause attachment. In addition, we found that we could not obtain significant improvement of the Vietnamese parsing by simply enlarging the training data. Among the three factors of word segmentation errors, POS tagging errors, and ambiguous constructions, although the first and second ones have significantly contributed to many parsing errors, the third one is the major problem that causes the low performance of Vietnamese parsing. Ambiguous constructions in Vietnamese appear in many forms, such as ambiguous POS sequences or ambiguous symbol sequences. They are caused by the characteristics of Vietnamese such as the lack of inflectional morphemes, post-head modifying lexical words, and dropping words. This research has also shown that although Vietnamese has many confusing constructions, these ambiguities can be tackled based on contextual information such as the words playing the roles as prefixes and suffixes, function words, fine-grained categorizations, head words of the phrases, main verbs of the clauses, etc.

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨

Summary of the results of the doctoral thesis screening

本博士論文は、ベトナム語の高精度な構文解析を実現するためのリソースとして、ベトナム語のツリーバンク（構文木がアノテートされたコーパス）を構築し、ベトナム語構文解析の問題点を明らかにすることを目的としている。

本論文は、全 5 章から構成される。第 1 章では、英語や中国語等におけるツリーバンク開発やそれを利用した研究について概説し、自然言語処理におけるツリーバンクの重要性を議論している。そして、ベトナム語の既存ツリーバンクはクオリティが不十分であること、ベトナム語の言語的特徴により高品質なツリーバンクを構築することが難しいことを挙げている。そこで、博士論文の貢献として、高品質なベトナム語ツリーバンクを構築するための手法を提案し約 20,000 文からなるツリーバンクを構築すること、そしてこのツリーバンクを利用して既存の構文解析モデルを評価し、ベトナム語構文解析の問題点を明らかにすることを主張している。

第 2 章では、本論文の背景として、ツリーバンク開発に関する方法論や既存研究、ツリーバンクを利用する研究として単語区切り、品詞タグ付け、構文解析の既存手法やこれらの技術の応用、ツリーバンク開発をサポートするための手法やツール、ツリーバンクを学習データとして利用する構文解析モデル、ベトナム語のツリーバンクや構文解析の既存研究について説明を行っている。

第 3 章では、高品質なベトナム語ツリーバンクを構築するための手法を提案している。まず、既存のベトナム語ツリーバンクおよび英語や中国語のツリーバンクのガイドラインを参照し、単語区切り、品詞、構文木のアノテーションガイドラインを構築した。特に、アノテーションが揺れるケースについて判断の根拠となるルールを設計した。また、高品質なアノテーションを行うためのアノテータの訓練・管理方法、およびアノテーションをサポートするツールを提案した。この結果、約 20,000 文からなるツリーバンクの構築を行った。ツリーバンクの品質を評価するため、アノテータ間一致度、アノテーションの精度、およびアノテータ内一致度を測定し、いずれも 90%以上を達成することが示された。さらに、既存のツリーバンクおよび本研究で構築したツリーバンクを用いて構文解析器を学習し、その精度を比較したところ、本研究のツリーバンクを用いた方が高い精度が得られることが示された。これにより、本研究により高品質なベトナム語ツリーバンクを構築したことが示された。

第 4 章では、このツリーバンクを用いて代表的な構文解析モデルを学習し、その精度を比較すること、さらに構文解析エラーを詳細に分析することによりベトナム語構文解析の問題点を明らかにすることを目的としている。まず、近年提案された構文解析モデルとして probabilistic context-free grammar (PCFG) に基づくもの、Conditional Random Fields (CRF) を用いるもの、Neural Network (NN) を用いるものを利用し、比較検討した。さらに、これらの構文解析器のエラーを分析することで、ツリーバンクのサイズは十分であること、ベトナム語構文解析特有のエラーとして VP attachment や NP attachment が頻出すること、品詞タグ付けのエラーが精度に影響するものの VP attachment や NP attachment の精度はほとんど変わらないこと、同じ品詞列に対して

(別紙様式 3)
(Separate Form 3)

異なる構文木が考えられる ambiguous construction がエラーの大きな部分を占めること、などが示された。これらは、ベトナム語構文解析のさらなる精度向上のための有用な知見である。

第 5 章では、以上の結果をまとめ、将来課題について議論している。

博士論文の内容については、提案手法、評価実験、エラー分析など、博士論文として十分なオリジナリティとクオリティがあるとの評価がなされた。本論文の内容は査読付きジャーナル「Language Resources and Evaluation」への採録が決定されている。以上のことから、全審査委員一致で本論文は学位授与に値するとの判断に至った。