

論文 / 著書情報  
Article / Book Information

題目(和文)	自然言語生成における文構造の制御に関する研究
Title(English)	
著者(和文)	丹羽彩奈
Author(English)	Ayana Niwa
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12462号, 授与年月日:2023年3月26日, 学位の種別:課程博士, 審査員:岡崎 直観,篠田 浩一,徳永 健伸,村田 剛志,井上 中順
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12462号, Conferred date:2023/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

## 論文要旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 知能情報	系 コース	申請学位 (専攻分野)： 博士 (工学) Academic Degree Requested Doctor of
学生氏名： Student's Name	丹羽彩奈		指導教員 (主)： Academic Supervisor(main) 岡崎直観
			指導教員 (副)： Academic Supervisor(sub)

### 要旨 (和文 2000 字程度)

Thesis Summary (approx.2000 Japanese Characters )

人間のように言葉を操れるコンピュータは、自然言語生成技術の発展によりさらに現実的なものとなっている。コンピュータが人間の文章の読み書きなどの言葉に関する技術を再現できれば、人間とコミュニケーションをとったり、作文の支援ができるようになる。そして実際に、対話システムや機械翻訳、自動校正など、多くの自然言語生成アプリケーションが人々の生活を仲介し、支援するようになった。この成功の背景にあるのは、深層学習技術の発展である。深層学習では、多層のニューラルネットワークからなるモデルを用いて入出力間に存在する複雑な関係を学習できる。これにより、多くの自然言語処理タスクの性能は飛躍的に向上し、商業化可能な水準に達した。一方で、深層学習におけるブラックボックス問題が近年問題視されている。これは、深層学習システムがなぜそのような出力をしたのかを人間が理解できない、つまり解釈性に欠けることを指す。これは最新の自然言語生成技術においても同様に見られ、出力文がどのような過程を経て生成されたのかを解釈することができない。この問題は、解釈性だけではなく、出力にある属性をもたせるための制御性の問題にまで波及する。特に自然言語生成においては、同じ内容を複数通りで表現可能であり、それによって文の難しさや効果に変化するため、制御性の高いシステムが強く求められる。しかし、ある属性を制御しようと思っても、それによって出力から制御の対象以外の情報が失われる可能性がある。そのためその想定外の挙動にも対処できるよう高度に生成を制御する技術が必要である。そこで本研究では、出力文にもたせたい文構造 (単語や句の並びや関係) で書かれたテキストを制約としてモデルに与え、生成される文の構造を制御する。この制御に用いるテキストを本研究ではテンプレートと呼ぶが、これはスロットをもつ雛形文であるハードテンプレートと、スロットをもたない見本となる用例であるソフトテンプレートのふたつに大別される。そしてそれらのテンプレートに対して、スロットに単語を穴埋めしたり、単語を削除・挿入することで文を生成する。このように、与えられたテンプレートを出発点として文生成を行うことで、生成される文の文構造を制御できるようになる。さらに、文生成の過程をテンプレートに対する一連の編集操作として人間が解釈できる。このテンプレートによって与える制約は、強いほど忠実にテンプレートの内容を守って文を生成できるため、強い構造制約を持つ言語表現に特に有効である。そこで、本研究ではまず強い制約を課せるハードテンプレートをモデルに与え、スロットを穴埋めすることで所望の文構造を確実にもった文を生成する解釈性の高い手法に取り組む。今回は、その強い構造制約をもつ文として対句と呼ばれる修辞技法のひとつを題材とし、その自動生成を行う。対句とは、「ひとに都合のいい空気は、地球に都合の悪い空気でした。」のように類似した文構造と対照的な意味を持つ文や句を並列させる技法であるが、対句を生成対象とした研究はなく、言語資源がない。そのため、本研究では対句文の自動生成を対句データセットの構築、ハードテンプレートの自動抽出と穴埋めの3つのサブタスクに取り組んだ。まず、高品質なデータ作成に向けて対句の判定基準を厳密に定義して対句データセットを構築した。次に自動抽出では、文中の対句に該当する範囲を計算量削減のための段階的な構造解析アルゴリズムによって同定し、その部分をスロットとすることで対句テンプレートを得る。そしてハードテンプレートへの穴埋めでは、対句の構造を認識した上で単語を穴埋めできる手法を提案した。結果として、定量的・定性的な評価により、提案したハードテンプレートに基づく手法は対句文を生成できることを示した。しかし、この手法は所望の文構造ごとにスロット位置を事前に同定する必要があり、また生成文の多様性に欠けるなどの問題がある。そこで、次はスロットをもたないソフトテンプレートを弱い構造制約としてモデルに与え、その文構造を参考にしつつ書き換えることによって文を生成するアプローチをとる。大量の文対から検索してきたソフトテンプレートに対し、複数トークンを一度に生成可能な非自己帰帰モデルを用いてトークンを挿入・削除することで文を生成する。提案手法は、多くのソフトテンプレートが取得可能な JRC-Acquis データセットを用いた機械翻訳タスクや WikiBio データセットを用いた Data-to-text タスクにおいて特に性能を向上させられることがわかった。また、分析結果によりソフトテンプレートによってモデル出力を弱く制御できていることが示唆された。

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。

Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

(博士課程)  
Doctoral Program

## 論文要旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 知能情報	系 コース	申請学位(専攻分野)： 博士 Academic Degree Requested Doctor of	( 工学 )
学生氏名： Student's Name	丹羽彩奈		指導教員(主)： Academic Supervisor(main)	岡崎直観
			指導教員(副)： Academic Supervisor(sub)	

要旨 (英文 300 語程度)

Thesis Summary (approx.300 English Words )

Research on controlling the black-box generation to obtain the desired text has gained increasing importance in natural language generation. In this research, the target of control is sentence structure, which is of great importance in constructing sentences.

First, we introduce a highly controllable filling-slot approach with the hard structural constraint. We focus on antitheses as the generation target. An antithesis is an expression that juxtaposes sentences or phrases that have similar sentence structures and contrasting meanings, such as “[There’re seconds.] But [there is no substitute.]” Their structural constraints are complex and require hard constraints. However, research on antithesis generation has not yet been reported. Thus, we start with (1) the construction of an antitheses corpus, followed by (2) template derivation and extraction of word pairs with contrasting meanings, and (3) template filling by adapting pre-trained BERT to antitheses generation. Consequently, quantitative and qualitative evaluations have demonstrated that the proposed hard template-based method can generate antitheses.

However, the hard template requires the pre-identified slot for each desired sentence structure, which costs and significantly reduces the diversity of the generated sentence. It also assumes that a limited number of hard templates can cover the desired sentence structure, limiting the applicable tasks.

Thus, the next step is to generate sentences by providing soft templates without slots as weak structural constraints. The proposed method generates sentences by inserting and deleting tokens for soft templates retrieved from many sentence pairs. This approach provides a high interpretability over the generation, as it is possible to know which tokens have been inserted or deleted at each editing stage by performing editing operations directly on the soft template. Experimental results demonstrate that the control of sentence structure by editing operations on soft templates is effective, e.g., in machine translation, and that the analysis results suggest that soft templates can weakly control model output.

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1 copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。

Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).