情報工学実験 II 10/18 論文紹介

・論文タイトル
自然言語とソースコード間の対訳コーパス向け Data Augmentation 手法の提案

・論文著者
秋信 有花、小原 百々雅、梶浦 照乃、倉光 君郎

・研究内容まとめ
自然言語とソースコード間の対訳コーパスの不足を解消するための Back-Translation による Data Augmentation 手法とその実装である DA ツール Multiese の提案を行っている。アノテーションと自然言語側の記法を導入することに加えプログラミング言語の性質を活用し自然言語表現の増強も行っている。

・著者の主張
ソースコードに対して DA 手法を適用し、ルールベースの記法を導入することで自然言語表現を多様化を実現した。加えて木構造に変換可能であるソースコードの性質を活用するという新しいアプローチを行った。

・興味を持った点
プログラムを実際に書いていると自分が実現したい機能からそれを行うためのソースコードをうまく書くことができないことが何度もあり、自然言語から直接ソースコードを生み出すことができればと思ったことが何度もあったので自然言語とソースコード間の対訳コーパスについてのこの論文に興味を持った。日本語と英語などの自然言語間での翻訳と異なりプログラミング言語で書かれたソースコードは木構造に変換可能で曖昧な表現を許さないという非対称性があるという点に論文を読んで面白いと感じた。また、プログラミング言語であるという性質を活かした Data Augmentation は異なるプログラミング言語間での翻訳など様々な場面で活用できそうだと思った。

・次に読むべきもの
Data Augmentation について目的や意義を理解することはできたが具体的にどのような手法があるのかについて知識が十分ではないと感じたので言語に対して行われる手法についてのものを読んどほうが良いと思った。back-translation についてはほとんど理解することができなかったので基礎的な説明を行っているものを中心に読むべきだと感じた。