ハングル文章の正誤判別の基礎研究

背景

・日本語にひらがながあるように韓国語にはハングルという文字体系がある。
・ハングルは他の文字体系より似た発音を持つ文字が多数あり、同音異義語やそれに近い単語が多く、発音は似ているが全く意味を持たない単語が生まれることもある。
・ある人はこれを悪用してハングルを意図的に損傷させ、翻訳機でも解釈できない文章を作っている。

実験

・背景のように、意図的に損傷された文章を修正するためにはハングルの発音と文脈を同時に理解する必要がある。
・しかし、入力された文章が正常な文章か間違った文章かを判断する段階では文脈や誤字脱字を分析することで十分。
•RNNを用いて正常な文章の文脈を学習。言い換えれば、ある文字列の次に出現する可能性が高い文字を学習。
・正常な文章の文脈を学習したContext Vectorは間違った文章では間違った文字に対して出現可能性が低いと答えるので、これを持って文章の正誤判別を行う。

今後

・ハングルの発音類似度を考慮し、文脈情報と合わせて損傷された文章の自動修正アルゴリズムを構築する。
・正常な文章データの収集を継続する。