特許情報に関する言語生成モデルを活用した知的財産創造手法の開発

Development of Intellectual Property Creation Method Using Language Generation Model on Patent Information

Shigeaki Onoda

Graduate School of Information Engineering, Toyama Prefectural University
t855005@st.pu-toyama.ac.jp

Wednesday., 7 8, 2019,
Toyama Prefectural Univ.

行ったこと

入力用データの作成

- キーワード抽出手法の選定
- 抽出ルール

作成手法

今回は東京大学中川研究室と横浜国立大学、森研究室が共同で作成した termextract というライブラリを用いて抽出する.
基本的なアルゴリズムは形態素解析結果を複合語に組み立て、２）その複合語（単語の場合もある）を重要度の高い順に返すもの
複合語により複雑な概念を表すことが多い専門用語を新規語も考慮してキーワードとして文章中から抽出すること可能となった

デモ

暫定的な解決法

以下リンクに web 版のデモがあるのでどのような単語を抽出できるか披露 http://gensen.dl.itc.u-tokyo.ac.jp/

抽出ルール

選定

上のライブラリでは英語にて 2 つの抽出法があるードで分割することで、専門用語を抽出する

1. 英文 POS Tagger の英文解析結果をもとに、専門用語を抽出する.
2. 英文を指定のストップワードで分割することで、専門用語を抽出する.

POS Tagger ：Python の NLTK モジュールの word_tokenize
この内ストップワードの方が未知語に対応し高速だが, Tagger を使ったほうが既知語に強いという特徴がある. いまのところ特許は形態素解析辞書に登録されてない用語が多いと感じたので 2 を選定.