Wikipedia と検索エンジンサジェストを用いた Web 検索語候補推薦システムの試作

宮下 竜大郎, 増田 英孝, 山田 剛一 (東京電機大学) 清田 陽司, 中川 裕志 (東京大学)
2010/09/15 国立情報学研究所 NLP若手の会 第5回シンポジウム

研究目的と背景

主要なウェブ検索エンジンにはサジェスト機能が用意されている。サジェスト機能は文字列の前方一致で候補を提示している。しかし、語の異表記やゆれに対する対応が不十分である。

表 1. Google サジェストによる検索語候補

「電機大」の検索語候補	「東大」の検索語候補
電機大学	東大
電機大学 2ch	東大 図書館
電機大学 偏差値	東大ノート
電機大学 理工学部	東大生協
電機大手	東大病院
電機大手 決算	東大寺
電機大手8社	東大寺学園
電機大手8社とは	東大和市
電機大手9社	東大和病院
電機大手9社とは	東大阪市

電機大は東京電機大学の異表記であり、東大は東京大学の異表記であるが、半数が東京電機大学または東京大学とは無関係なサジェストである。このように、利用者が意図しない検索語候補が提示されることがある。

本研究では Wikipedia の情報と検索エンジンサジェストを組み合わせることによって異表記やゆれに対応し、提示する検索語候補を分類する検索語候補推薦システムの試作を行った。

今回のシステムでは Wikipedia のリダイレクト情報と記事タイトルを用いている。

情報源としての Wikipedia

リダイレクト機能

Wikipedia には原則として正式名称でページを構築する決まりがある。しかし、リダイレクトページを構築しておくことによって、略称や別名で記事を検索した際にも正式名称を持つ記事へ自動的に転送させることができる。

このリダイレクト情報を用いることで、検索語の曖昧性を回避させることができる。例えば記事「東京電機大学」にはリダイレクトページとして「電大」「電機大」が設定されている。この情報を用いることで、電大または電機大と入力した場合においても、東京電機大学のサジェストと提示することができる。

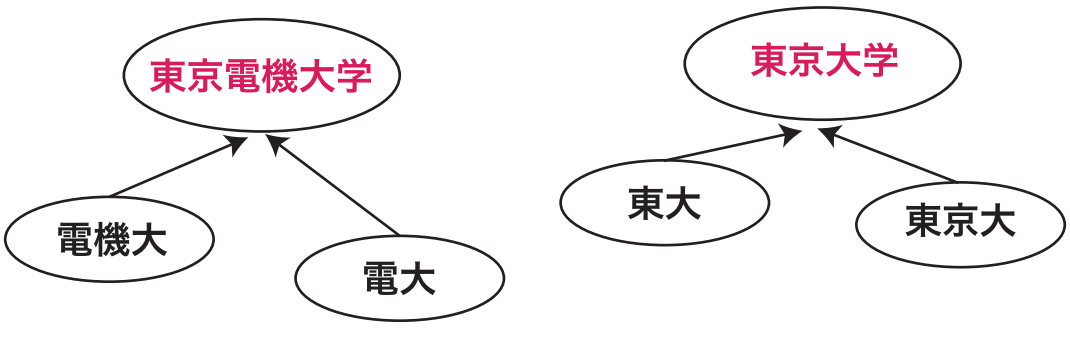

図 1. リダイレクト構造の例

提案手法とシステム

今回実装したシステムは検索語の曖昧さを回避と検索語候補の分類を目的としている。流れは以下の通りである。

1. キーワードの入力
2. キーワードを用いて Wikipedia のリダイレクト情報を参照
3. リダイレクト有) 入力語とリダイレクト語からサジェストを取得
   リダイレクト無) 入力語からサジェストを取得
4. サジェストの取得に用いた語から前方一致検索で Wikipedia の記事タイトルを検索し取得する
5. 記事タイトルがサジェストに含まれているかどうかを確認する
6. 記事タイトルと一致するものを 1 グループとしてグループ化
7. 入力語前後に空白スペースが無いものに (固有名詞) タグを付加する

検索エンジンサジェスト		Wikipediaの記事タイトル	
入力語	リダイレクト語	入力語	リダイレクト語
電機大 秋葉原	東京電機大学 2ch	電機大	東京電機大学
電機大学 2ch	東京電機大学 事務部	電機子	東京電機大学出版局
…	…	…	…
電機大手	東京電機大学出版局	電縁	東北弁

-----[東京電機大学]-----
- 東京電機大学
- 東京電機大学 2ch
- 東京電機大学 オープンキャンパス
- 東京電機大学 メディアセンター
- 東京電機大学 偏差値
- 東京電機大学 理工学部

〜

-----[電機大]-----
- 東京電機大(固有名詞)
- 秋葉原 通り魔 電機大
- 秋葉原 電機大
- 通り魔 電機大
- 電機大 スレ
- 電機大 傷害

〜

- 電機大学(固有名詞)
- 電機大学 2ch(固有名詞)
- 電機大学 偏差値(固有名詞)
- 電機大学 理工学部(固有名詞)
- 電機大学 飛び降り(固有名詞)
- 電機大手(固有名詞)
- 電機大手 決算(固有名詞)
- 電機大手8社(固有名詞)
- 電機大手8社とは(固有名詞)
- 電機大手9社(固有名詞)
- 電機大手9社とは(固有名詞)

電機大手として分類されている

図 2. 出力結果例

-----[macbook air]-----
- macbook air

-----[macbook pro]-----
- macbook pro
- macbook pro 13
- macbook pro 2010
- macbook pro ssd
- macbook pro レビュー
- macbook pro 価格
- macbook pro 新型

-----[macbook]-----
- macbook
- macbook coolなデザインのk.collection
- macbook ssd
- macbook ケース

〜

タイトル情報によって製品名のような場合でも分類ができている

図 3. 出力結果例 2

終わりに

検索語から Wikipedia のリダイレクト情報を参照することで異表記やゆれを吸収し、得たサジェストを Wikipedia の記事タイトルとマッチングさせることで問題の解決を図った。Wikipedia のカテゴリ情報の利用やシステム、手法の評価に関して、今後検討する必要がある。