意味的類似度を利用した
日本語クエリ書き換えのための
統一的アプローチ

萩原 正人(†) 鈴木 久美 (‡)
(†) 名古屋大学
(‡) Microsoft Research

背景

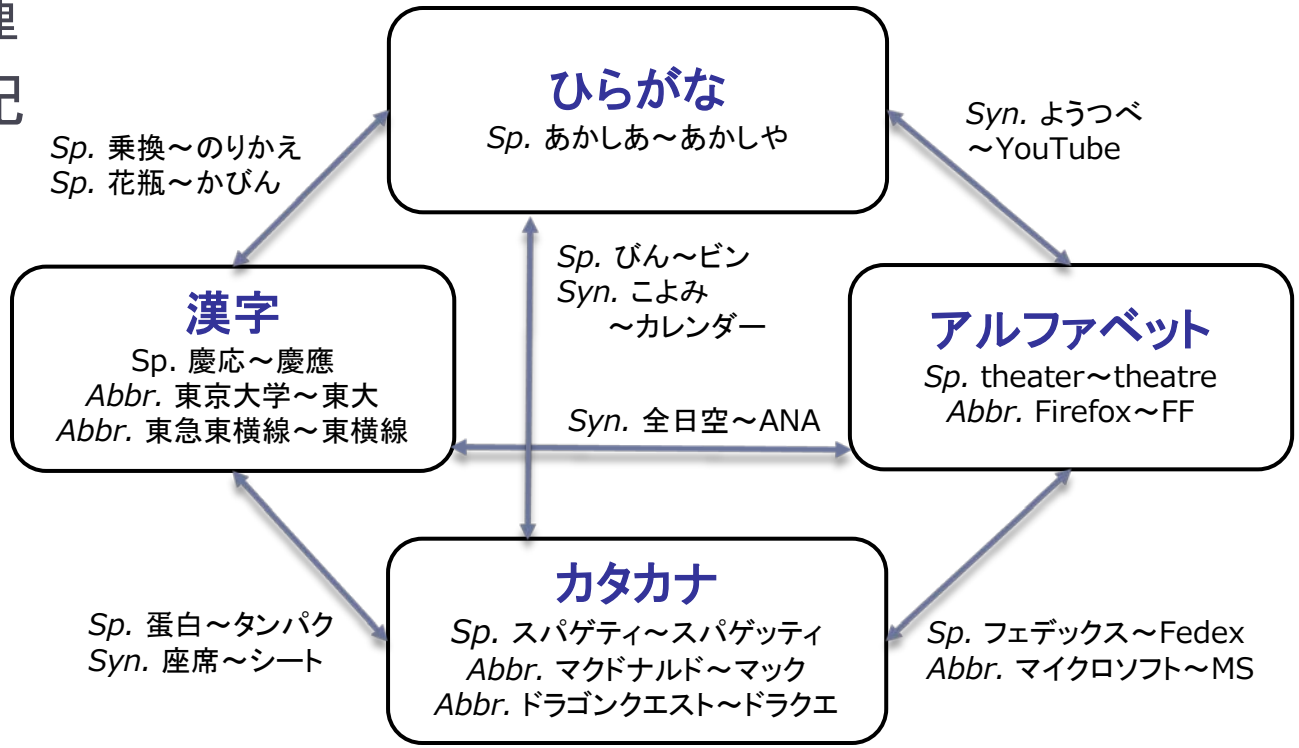
- クエリ訂正
  - ロバストな検索エンジンを実現する上で必須の技術
  - クエリの10%以上は誤りを含む (Cucerzan and Brill 2004)

“You Tube”のミススペル (Microsoft Live Searchで入力された実際の例)

Tou Tube	u-tubu	you tyub	youtyube
U-チューブ	yiutube	you tyubu	youube
YOIUTUBE	yo tube	you utbe	youutube
YOU TUBE	yoiutube	yoube	youytube
YOUTBE	yotube	youbube	yoy tube
YOUTUBE	yotuube	youitube	ypoutube
YOUTUBU	you btube	youtbe	yutube
YOUUBE	you chube	youtibe	ЮOチューブ
You Tube	you tebe	youtobe	ゆうチューブ
You Tubu	you tobe	youtube	ゆーちゅーぶ
You tube	you trube	youtube]	ウーチューブ
YouTube	you tube	youtubu	ユウチュウブ
YouTube	you tube[	youtubwe	ユウチューブ
Youtube	you tubr	youtueb	ユーチュウブ
Youtube	you tubue	youtuge	ユーチュウーブ
Youtubu	you tubw	youtuube	ユーチュー
oyutube	you tuibe	youtuybe	ユーチューブ
tou tube	you tybe	youtybe	ユーチューヴ
			ユーテューブ

問題

- 日本語固有の問題
  - 多様な文字種
  - 翻字の異表記



- スペル訂正以外の書き換え
  - 異表記，略称，同義語

目的

- 特定の文字種や異表記に限定しない統一的な日本語クエリ書き換えモデルを提案
  - 表記の類似度と意味的な類似度に基づいた英語のクエリ訂正手法に基づく
  - 単純なスペル訂正だけでなく，異表記の統一（スパゲッティ～スパゲティ），略称（MS～マイクロソフト），同義語（座席～シート）なども対象
  - 意味的類似度に基づくカーネル法によって，書き換え候補間の類似度をロバストに推定