# 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（名古屋大学） 鈴木 久美（Microsoft Research）

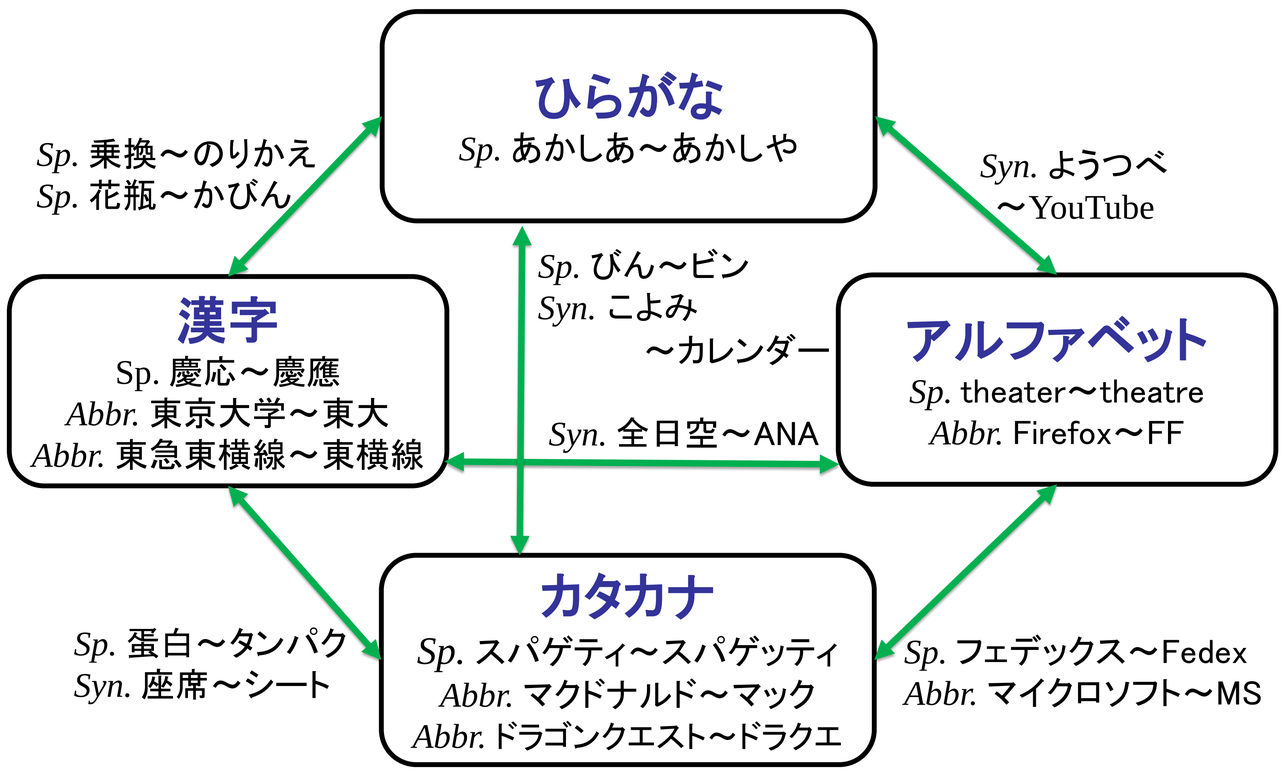
## 背景: クエリ書き換えの必要性

- 10%以上のWeb検索クエリは誤りを含む
- 英語以外の言語では未着手
  - 英語のクエリ訂正 (Cuceran and Brill 04), (Li et al. 06), ...
- 日本語特有の問題
  - 文字種: e.g. たんぱくしつ，タンパク質，蛋白質，...
  - 翻字の異表記: e.g. スパゲティ，スパゲッティー

## 目的: クエリ書き換えの統一的フレームワーク

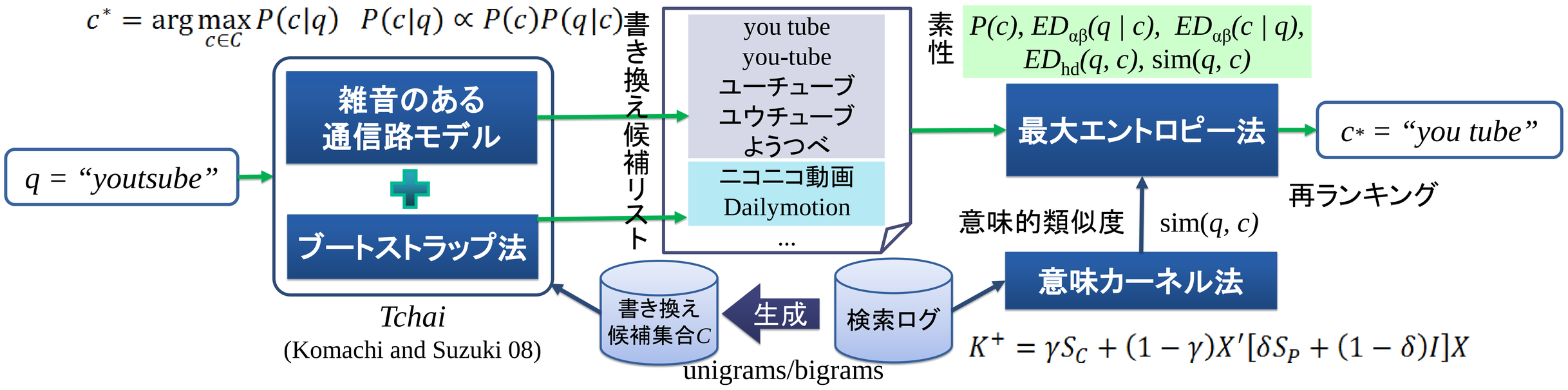
- 特定の文字種や表記の類似度に限定しない
- スペル訂正の他に，異表記，略称，同義語も対象
- 意味的類似度に基づくカーネル法の利用

## 日本語における文字種と異表記




## クエリ書き換えモデル

$$c^* = \arg\max_{c \in C} P(c|q) \quad P(c|q) \propto P(c)P(q|c)$$

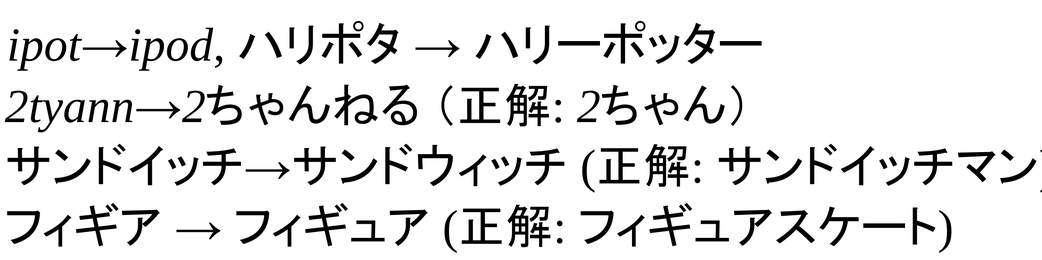



### 雑音のある通信路モデル

- 言語モデル $P(c)$ ... 検索ログ中の相対頻度
- 英語-仮名(ローマ字)の翻字モデル $ED_{\alpha\beta}(q \mid c)$
  - 編集距離の一般化 (α→β) (Brill and Moore 00)
    Wikipediaの対応付きタイトル(59Kペア)から学習
- 仮名(ローマ字)間の異表記モデル $ED_{hd}(q, c)$
  - 繰り返し(aa→a, a→aa)のコストを0にした編集距離
- 漢字の読みモデル (予定)
  - (単漢字, 読み)のbigram言語モデル

### 意味カーネル法

- 検索ログ中の文脈の分布類似度 (Li et. al 06)
- 候補-文脈の共起グラフカネール (Kandora et al. 02)
  - von Neumann Kernel, Exponential Kernel
  - 文脈の相関を考慮可能
- 補完カーネルにより文脈のスパースネスを解消




## 実験

### 訓練データの作成

- 書き換えペア候補を提示し，効率的に作成
  - 従来手法: *gogle* を提示 → ユーザー意図が不明
  - 提案手法: (“gogle”, “google”) → OK/NGを判断
- ペア候補はセッションログから自動生成
  - 同じユーザーが3分以内に$q_1$と$q_2$を入力
  - $q_1$のクリック数 = 0 かつ $q_2$のクリック数 > 0
  - 対数尤度比 LLR($q_1$, $q_2$ ) が閾値以上 (Jones et al. 06)
- 評価者間で高い一致率 (95.7%)
- データセット
  - Live Search 検索ログ2ヶ月分 (約100万異なりクエリ)
  - Train: 4,618ペア, Dev.: 628ペア, Test: 1,243ペア

### 結果: 各モデルの性能比較

| | モデル | 正解率 | 再現率 | 精度 |
|---|---|---|---|---|
| 雑音のある通信路 | SC | 71.12 | 63.79 | 54.76 |
| 類似度なし | ME-NoSim | 74.58 | 70.20 | 60.77 |
| コサイン類似度 | ME-Cos | 74.18 | **71.18** | 59.10 |
| カーネル法 | ME-vN | 74.34 | **71.18** | 60.33 |
| | ME-Exp | 73.61 | 70.20 | 59.01 |
| 補完カーネル法 | ME-vN+ | 75.06 | 69.70 | 61.52 |
| | ME-Exp+ | **75.14** | 68.23 | **61.56** |

- OK: *ipot*→*ipod*, ハリポタ → ハリーポッター
- NG: *2tyann*→*2*ちゃんねる（正解: *2*ちゃん）
- NG: サンドイッチ→サンドウィッチ (正解: サンドイッチマン)
- NG: フィギア → フィギュア (正解: フィギュアスケート)
  - 限定的な正解