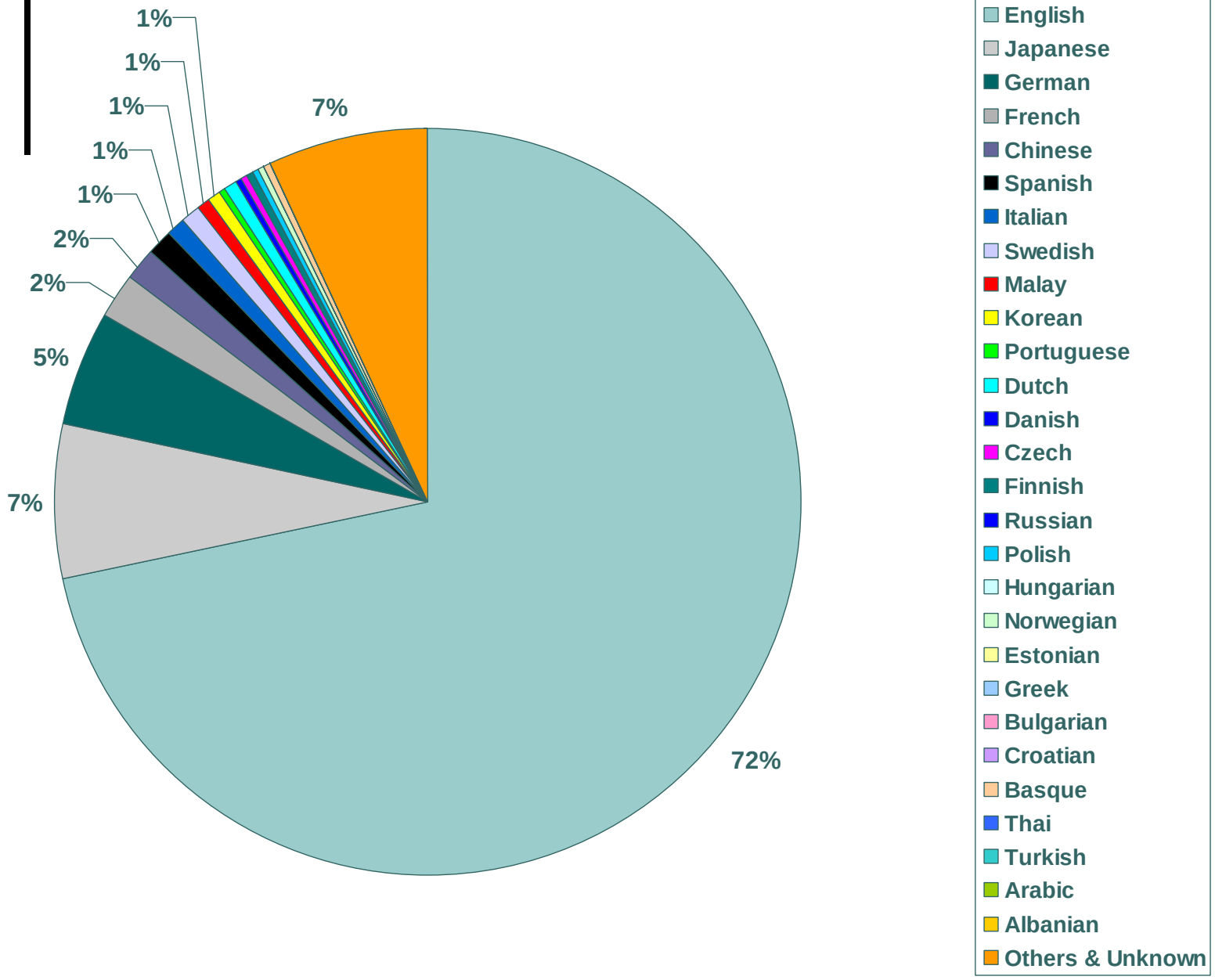
ディレクトリ型検索エンジンのカテゴリ間対応付けによる言語横断検索

前田 亮 (科学技術振興事業団CREST)
木村 文則, 吉川 正俊, 植村 俊亮 (奈良先端科学技術大学院大学)

研究の背景

- 通常のWeb検索（日本語による問合せ）
  - 日本語だけではWebの一部しか検索できない
- 検索要求によっては，他の言語も探したい
  - ある国のニュースは，その国のニュースサイトのほうが情報が豊富

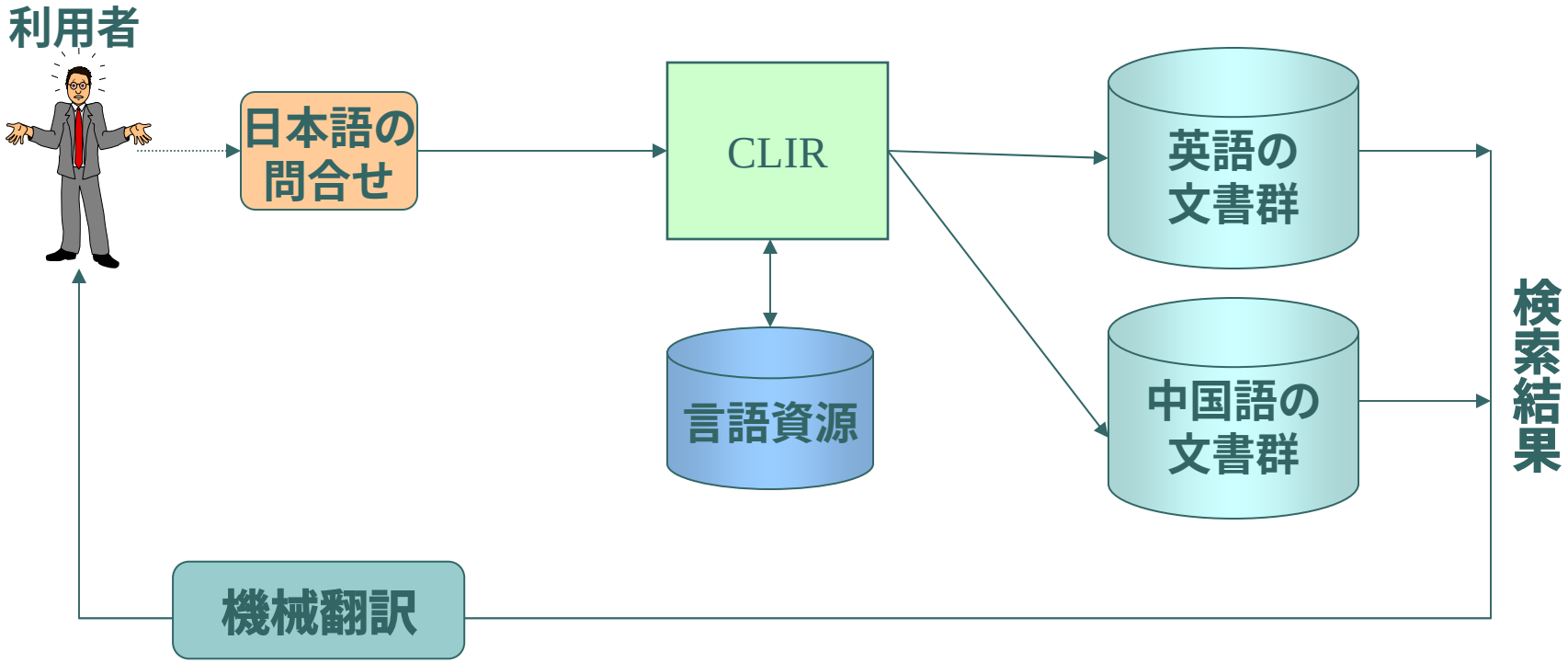
Webに用いられている言語



Source: Jack Xu, Excite@Home, 1999

言語横断情報検索
(Cross-Language IR: CLIR)

- ある言語で書かれた文書群を，別の言語による問合せで検索する



言語横断検索へのアプローチ

- 検索対象文書を翻訳
  - 既存の機械翻訳システムが使用可能
  - Webのように，大規模で更新が頻繁な文書群に対しては非現実的
- 利用者の問合せを翻訳
  - 検索対象文書群の規模・更新頻度は問題にならない
  - 翻訳された問合せは，既存の検索エンジンにそのまま適用可能

問合せ翻訳手法の問題点

- 対訳辞書には訳語候補が複数存在する（訳語の曖昧性）
  - 例：bank→銀行，堤防，土手，川岸 …
- 問合せをまず対訳辞書で翻訳し，曖昧性解消にコーパスを用いる手法が主流
  - 共起頻度などの情報を利用
    - “economy bank ” → “経済 銀行”
    - “river bank ” → “川 堤防”

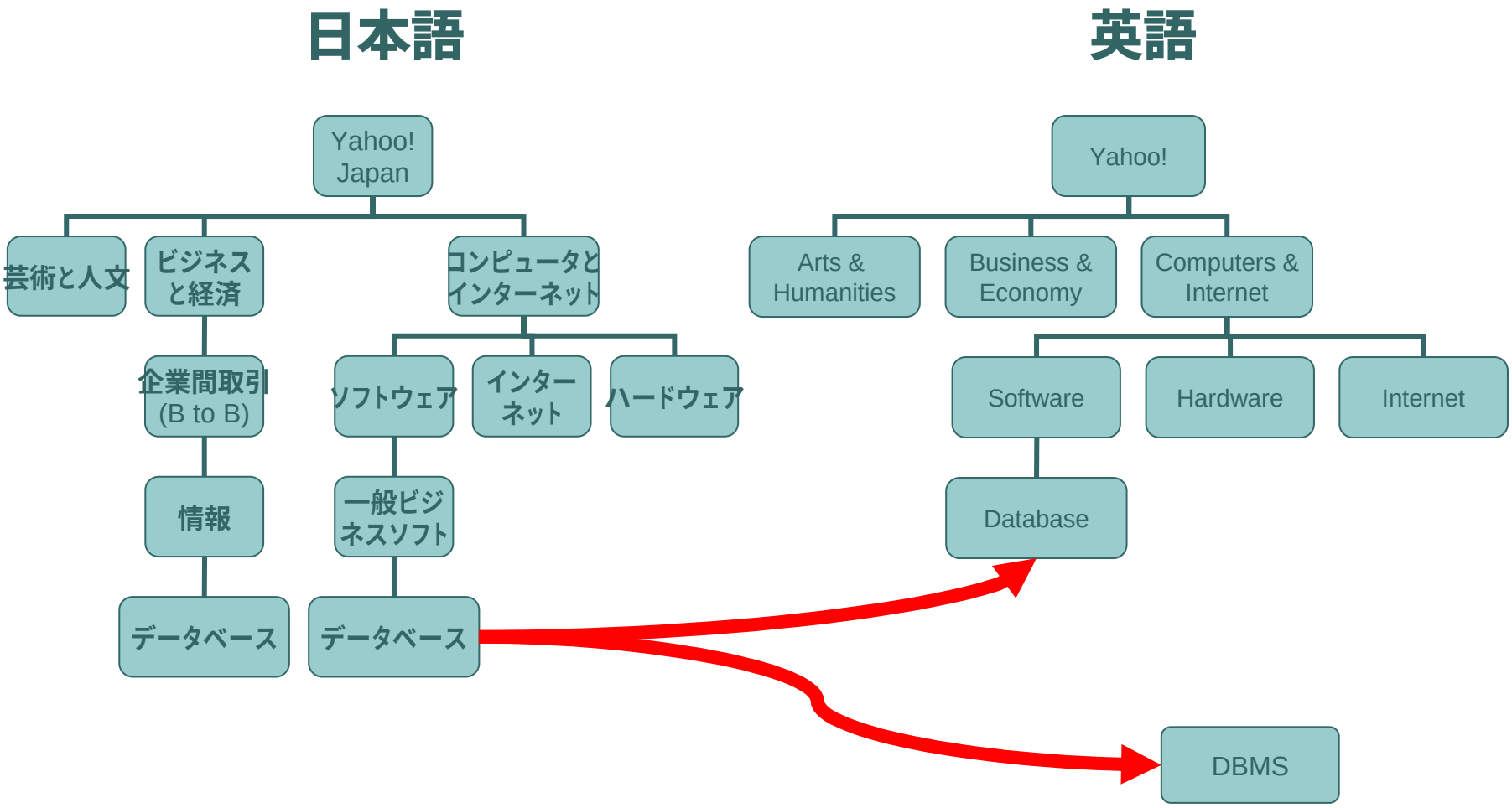
コーパスを用いる手法の問題点

- 分野に対する依存性
  - 分野の違うコーパスでは，曖昧性解消が難しい
  - Webは様々な分野の文書が混在
    - すべての分野を網羅したコーパスは入手困難なため，曖昧性解消が困難
- 規模の問題
  - 多様な分野を網羅した大規模なコーパスは入手困難

提案手法 (前処理)

- 複数言語で類似のカテゴリ構造を持つディレクトリ型Web検索エンジン
  - Yahoo!の各国語版を利用
- 前処理：異言語カテゴリ間の対応付け
  1. 各カテゴリの特徴語を抽出
  2. 特徴語を翻訳
  3. 異言語の適合カテゴリを選択

異言語カテゴリ間の対応付け



特徴語の抽出手法

- あるカテゴリ中で，より多くの文書に出現している単語ほど重要と仮定

$$df(t^c) = \frac{\sum_{d=1}^{N} w_{t^c}^{d}}{N}$$

N:そのカテゴリに属する文書数
W^d_{tc}:カテゴリcに属する文書dにおける索引語tの重み:

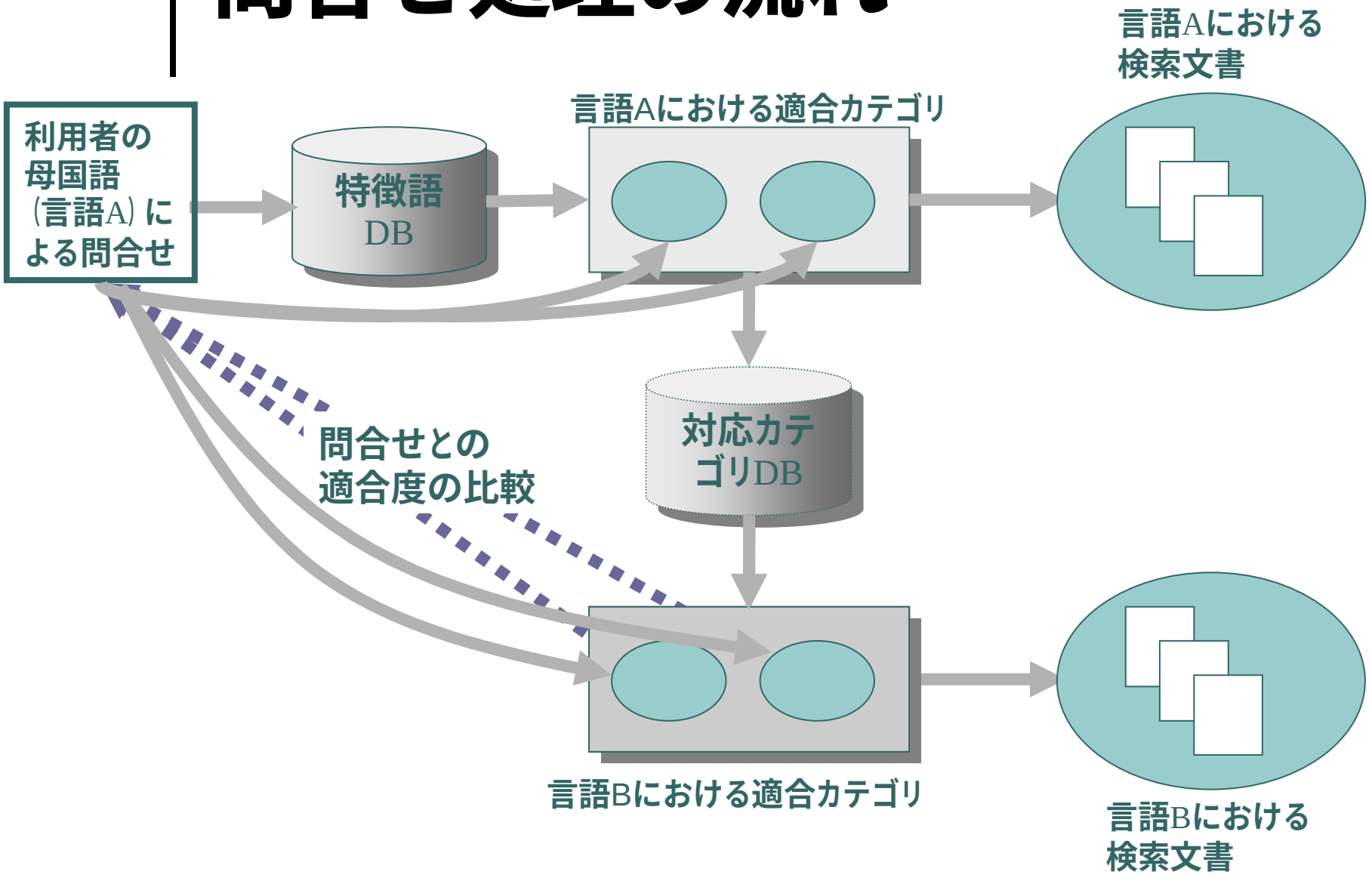
- 重みの大きいものからn語（もしくはある閾値以上となるもの）をそのカテゴリの特徴語とする

問合せに対する処理

- 検索対象言語の対応カテゴリのみ検索
  1. 問合せの同言語適合カテゴリの選択
  2. 異言語適合カテゴリの決定
  3. 適合カテゴリ内の文書を検索

訳語の曖昧性解消と検索性能の向上

問合せ処理の流れ


利用者の
母国語
(言語A)に
よる問合せ
特徴語
DB
言語Aにおける適合カテゴリ
言語Aにおける
検索文書
問合せとの
適合度の比較
対応カテ
ゴリDB
言語Bにおける適合カテゴリ
言語Bにおける
検索文書

予備実験

- **カテゴリの特徴語の抽出**
  - Yahoo! Japan：**コンピュータとインターネット – ソフトウェア – 一般ビジネスソフト – データベース**
  - Yahoo!: Computers & Internet – Software – Databases
- **各カテゴリの登録Web文書の1リンク先まで取得**
  - **日本語34件 (約330KB)，英語44件 (約** 634KB)

実験結果

	Yahoo! Japan	Yahoo!
1	ページ	mysql
2	して	support
3	する	training
4	し	search
5	re:	server
6	Home	partners
7	さ	cc
8	こと	database
9	Page	mailing
10	日	consulting

実験結果の考察

- 日本語・英語ともに不適切な単語が多く含まれていた
  - 日本語の形態素解析の問題（「して」「する」「こと」など）
  - 1リンク先だけでは充分なデータ量が得られなかった
  - 特徴語の抽出手法の問題
    - 出現頻度だけではうまく抽出できない

まとめ

- 複数言語で類似の構造を持つWebディレクトリを言語横断検索に利用する手法を提案
  - Webディレクトリの言語資源としての活用
  - 他に対訳辞書しか必要としない
  - 雑多の分野が含まれるWeb文書に対する言語横断検索に有効

今後の課題

- 特徴語の抽出手法の再検討
  - 語の接続関係，品詞解析，etc…
- カテゴリ間対応付けおよび検索の実験
- カテゴリの階層構造の利用
- テストコレクションを用いた評価実験