CROSS LANGUAGE INFORMATION RETRIEVAL (CLIR)

Pengantar Temu-Kembali Informasi

Kuliah #12

6 Maret 2009

Latar Belakang

- Naiknya kebutuhan untuk mengakses informasi tanpa halangan bahasa atau budaya, yang berarti ada permintaan yang kuat untuk dapat:
  - Menemukan informasi yang ditulis dalam bahasa asing
  - Membaca dan menginterpretasikan informasi dan menggabungkannya dengan informasi pada bahasa-bahasa lain
- Kebutuhan adanya Multilingual Information Access

Monolingual vs CLIR

- Monolingual IR
  - Memperoleh dokumen yang bahasanya sama dengan query
- CLIR
  - Memperoleh dokumen yang bahasanya berbeda dengan bahasa yang ada pada query

Pengertian

- Cross-language
  - Cross-lingual, cross-linguistic, translingual
- Dokumen Multilingual
  - Dokumen berisi lebih dari satu bahasa
- Koleksi Multilingual
  - Koleksi dokumen dalam bahasa yang berbeda-beda
- Multilingual system
  - Dapat memperoleh dokumen dari suatu koleksi multilingual
- Cross-language system
  - Query dalam bahasa yang satu, cari dokumen dalam bahasa lain
- Translingual system
  - Query dapat menemukan dokumen dalam bahasa apapun

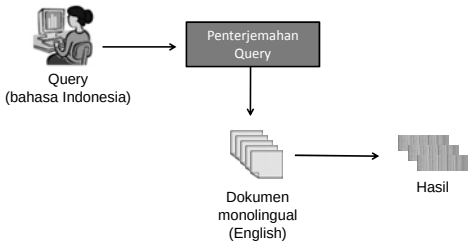
Disain Sistem

- Apa yang perlu di-indeks?
  - Free text atau controlled vocabulary
- Apa yang perlu diterjemahkan?
  - Query atau dokumen
- Di mana kita bisa mendapatkan resources untuk menerjemahkan?
  - Kamus, ontologi, training corpus

Dokumen vs Query

- Penerjemahan Dokumen
  - Menerjemahkan dokumen ke bahasa dari query
  - Tidak praktis. Prosesnya lambat, walaupun hanya perlu menterjemahkan sekali untuk setiap dokumen.
- Penerjemahan Query
  - Menerjemahkan query ke bahasa dari dokumen
  - Efisien untuk query yang pendek

Contoh



Metode Penterjemahan

- Mesin Penterjemah
- Kamus Dwibahasa
- Korpus Paralel
- Transitif

Mesin Penterjemah

- Belum tersedia pada banyak bahasa
- Contoh:
  - SYSTRAN, LOGOS, Langenscheidt tersedia dalam bahasa Jerman, Perancis, Inggris, dan Spanyol
  - ASTRANSAC (Jepang-Inggris)
  - Toggletext (http://www.toggletext.com), BPPT, dan Transtools (Indonesia-English)
- Keterbatasan:
  - Dasar dari mesin penerjemah adalah aturan linguistik sehingga hasilnya akan baik jika query ditulis dalam kalimat sesuai dengan tata bahasa yang baik.
  - Seringkali tidak dapat menerjemahkan kata gabungan dan *proper nouns*.

Kamus Dwibahasa

- Tersedia secara luas, menghasilkan daftar kata dwibahasa.
- Contoh kamus dwibahasa:
  - Collins
  - Kamus gratis : http://www.freedict.com
  - Babylon : http://www.babylon.com
  - Linguistic Data Consorsium
  - EuroWordNet

Contoh Query

- Penterjemahan per kata dalam query
- Contoh (English Query): The effects of chocolate on health
- Query Indonesia: Pengaruh permen coklat pada kesehatan
- Diterjemahkan dengan kamus Indonesian-English:

  influence hard candy brown chocolate cocoa health

Frase ?

- Contoh (English Query): Reasons for controversy surrounding Waldheim's World War II action
- Diterjemahkan dengan kamus Indonesian-English: controversy measure action step waldheim world kingdom war battle II

Frase ?

- Tidak dapat menerjemahkan frase jika kamus tidak berisi frase tersebut
- Penggunaan kata yang berbeda di bahasa yang lain
- Contoh:
  - acupuncture (satu kata)
  - tusuk jarum (dua kata)
    - Terjemahan dari kamus : tusuk – puncture; jarum – a pin; a stick, skewer, sewing or hypodermic needle; pin; hand of clock, pointer

Pemecahan Masalah

- Menggunakan POS-taggers; kamus khusus; pilih definisi yang terbaik atau yang pertama saja; feedback dari user
- Teknik memilih kata terjemahan yang tepat berdasarkan pada analisa statistik. Kata yang nilainya paling tinggi yang dipilih sebagai kata terjemahan.
- Contoh:
  - Peudo relevance feedback
  - Local context analysis
  - Term Similarity
  - Probabilistic

Identifikasi Frase

- Semantik (misalnya : yang muncul di kamus)
- Sintaktik (misalnya : diperoleh sebagai frase kata benda)
- Co-occurrence (kata yang sering muncul bersama)

QE via Penterjemahan

- Pre-Translation Query Expansion
  - Menambahkan kata-kata pada query sebelum diterjemahkan
  - Memperbaiki query
- Post-Translation Query Expansion
  - Menambahkan kata-kata pada query sesudah diterjemahkan
  - Mengurangi kesalahan penterjemahan
- Kombinasi Pre- & Post-Translation QE
  - Menambahkan kata-kata pada query sebelum dan sesudah diterjemahkan

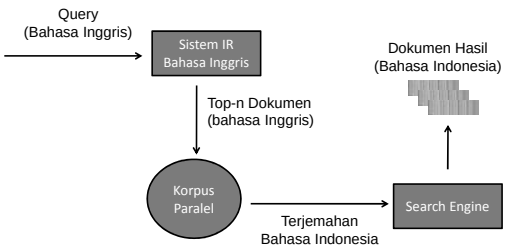
Korpus Paralel

- Korpus paralel: koleksi berisi dokumen yang sama dalam beberapa bahasa
  - Pasangan dokumen
  - Pasangan kalimat
  - Pasangan kata
- Comparable corpora (korpus yang sebanding)
  - Koleksi berisi dokumen dengan topik, waktu yang sama. Misalnya Kantor berita ABC melaporkan dalam bahasa Inggris dan Indonesia
  - Pasangan koleksi
  - Pasangan dokumen

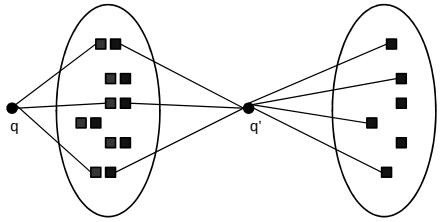
Menterjemahkan Query

- Pasangkan dokumen yang berkaitan melalui deskriptor (tanggal, kata kunci, kata benda nama)
- Buat leksikon dari co-occurence
- Kata-kata pada bahasa lain yang menunjuk pada topik yang sama akan muncul sama-sama pada tiap dokumen
- Gunakan hubungan pada query yang diterjemahkan secara semu (Pseudo-translation)

Pseudo-Translation



Pseudo-Translation



Pasangan Kalimat

- Mudah dibuat dari dokumen yang dipasangkan
  - Cocokkan pola dari panjang kalimat yang relatif
- Pasangkan kata-kata menggunakan statistik co-occurrence
  - Seberapa sering suatu pasangan kata muncul pada pasangan kalimat?
  - Bobotnya tergantung pada posisi relatif pada kalimat
  - Buang pasangan kata yang munculnya tidak sering
- Berguna untuk penterjemahan query
  - Hasilnya baik bila domainnya sama
- Belum secara langsung digunakan untuk retrieval yang efektif

Terjemahan Transitif

- Jika sumber daya penterjemahan terbatas antara dua bahasa, maka bisa dilakukan penterjemahan melalui bahasa lain