KULIAH 8

CROSS LANGUAGE INFORMATION RETRIEVAL

Bab 4: Grossman

Masalah

- 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 & menginterpretasikan informasi dan menggabungkannya dengan informasi pada bahasa-bahasa lain
- Kebutuhan adanya Multilingual Information Access

Mengapa CLIR Penting?

- Internasionalisasi
  - Negara-negara Multilingual (Switzerland, Canada)
  - Area kerja sama ekonomi (EU, EFTA, NAFTA)
- Globalisasi ekonomi
  - Perusahaan multinasional
  - Pegawai berbicara dalam berbagai bahasa
  - Pelanggan berbicara dalam berbagai bahasa
  - dokumen memerlukan akses dalam berbagai bahasa

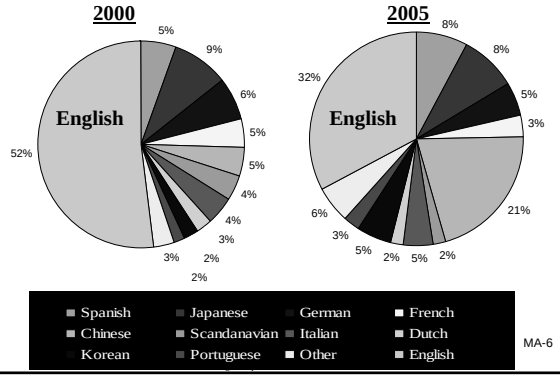
Internet

- Internet tidak lagi monolingual dan isinya yang bukan bahasa Inggris berkembang sangat cepat
- Perubahan profil pemakai sangat besar
  - Awalnya dari akademik, lalu digunakan secara luas pada komersial, hiburan, pendidikan, dll.

Perubahan pada Internet

- Pada 2005, 78% pemakai internet adalah orang-orang yang bahasanya bukan Inggris
- Jumlah pemakai internet akan naik dari 171 juta menjadi 345 juta pada 2005
- Karena itu… ada 270 juta pemakai Internet yang bahasanya bukan Inggris pada 2005.

78% Pemakai Internet Bahasanya Bukan Inggris pada 2005



Cross-Language IR

- Monolingual IR
  - Memperoleh dokumen yang bahasanya sama dengan query
- CLIR
  - Memperoleh dokumen yang bahasanya berbeda yang bahasa pada query
- Bila *user* dapat membaca dalam beberapa bahasa
  - Menghilangkan berbagai query
  - Query ditulis dalam bahasa yang paling dikuasai

Perbendaharaan Kata

- **Cross-language**
  - Cross-lingual, cross-linguistic, translingual
- **Dokumen Multilingual**
  - Dokumen berisi lebih dari satu bahasa
- **Koleksi Multilingual**
  - Koleksi dokumen dalam bahasa-bahasa yang berbeda
- **Multilingual system**
  - Dapat memperoleh dokumen dari suatu koleksi multilingual

Perbendaharaan Kata

- *Multilingual System*
  - Digunakan untuk menjelaskan *cross-language systems*
    - Query dalam bahasa Inggris, cari dokumen dalam bahasa Inggris atau Perancis
    - Query dalam bahasa Perancis, cari dokumen dalam bahasa Inggris atau Perancis
  - Juga digunakan untuk sistem pasangan monolingual
    - Query dalam bahasa Inggris, cari dokumen dalam bahasa Inggris
    - Query dalam bahasa Perancis, cari dokumen dalam bahasa Perancis

Perbendaharaan Kata

- *Cross-language system*
  - Query dalam bahasa yang satu, cari dokumen dalam bahasa *lain (another language)*
- *Translingual system*
  - Query dapat menemukan dokumen dalam bahasa *apapun (any language)*

Keputusan Perancangan

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

Penerjemahan Dokumen vs Query

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

Cross-Language Information Retrieval

Teknik Perbendaharaan Kata Terkontrol

(*Controlled Vocabulary*)

Bagaimana Cara Kerja *Controlled Vocabulary*

- *Thesaurus Design*
  - Design suatu struktur *knowledge* untuk domain
  - Beri suatu "*descriptor*" unik untuk tiap konsep
- Mengindeks Dokumen
  - Baca dokumen, beri *descriptor* yang sesuai
- *Retrieval*
  - Pilih *descriptor* yang diinginkan, gunakan *exact match retrieval*
- Tiga teknik pembuatan *multilingual thesaurus*
  - Buat dari awal
  - Terjemahkan thesaurus yang ada
  - Gabungkan thesaurus monolingual

Kelebihan dari *Controlled Vocabulary*

- Pengindeksan berdasarkan konsep yang berkualitas tinggi
  - *Descriptor* tidak perlu muncul pada dokumen
- Pelacakan yang dibimbing oleh *knowledge*
- Efektifitas cross-language sangat baik
  - Hingga 100% dari efektifitas monolingual
- Hasil retrieval mudah dimengerti
- Implementasinya efisien

Keterbatasan

- Biaya pembuatan sangat besar
  - Design struktur *knowledge*, indeks setiap dokumen
- Biaya pemeliharaan sangat besar
  - Pengindeksan dokumen, *vocabular*y dan perubahan konsep
- Sukar menggunakannya
  - Pilihan *vocabulary*
- Lingkupnya terbatas
  - Domain harus dipilih pada saat perancangan

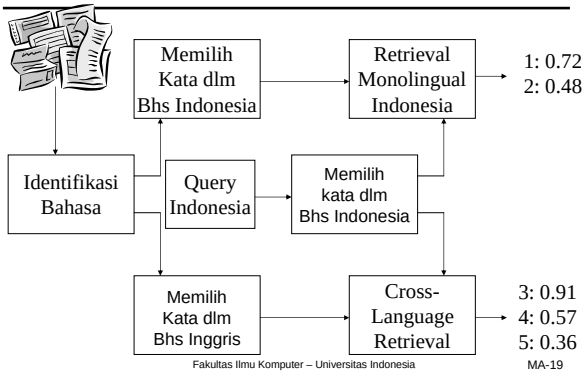
Cross-Language Information Retrieval

Teknik Berdasarkan Knowledge
Untuk Pelacakan *Free Text*

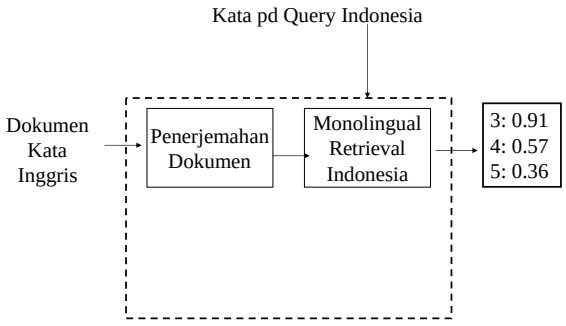
Struktur *Knowledge* untuk IR

- Ontologi
  - Representasi dari konsep dan hubungannya
- Thesaurus
  - Ontologi khusus untuk retrieval
- Leksikon Dwibahasa
  - Ontologi khusus untuk mesin penerjemah
- Kamus Dwibahasa
  - Ontologi khusus untuk penerjemahan yang dilakukan manusia

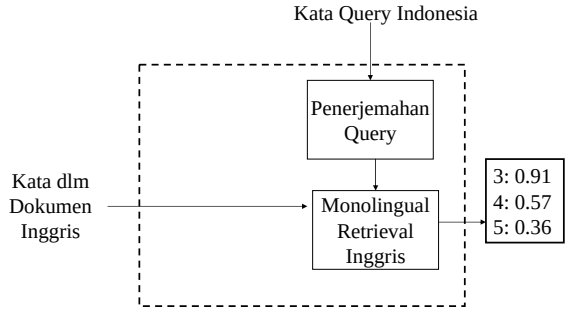
Arsitektur *Translingual Retrieval*



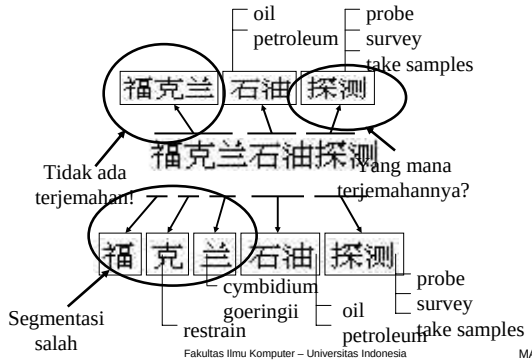
Retrieval dengan Bahasa sesuai Query



Retrieval dengan Bahasa sesuai Dokumen



Tantangan di CLIR : Segmentasi



Cara Penerjemahan

- Mesin Penerjemah
- Kamus Dwibahasa
- Korpus Paralel

Penggunaan Mesin Penerjemah

- Berdasarkan pada NLP
- Belum tersedia pada banyak bahasa
- Contoh mesin penerjemah yang tersedia di Internet:
  - *SYSTRAN, LOGOS, Langenscheidt* tersedia dalam bahasa Jerman, Perancis, Inggris, dan Spanyol
  - ASTRANSAC (Jepang-Inggris)
  - Toggletext (http://www.toggletext.com), BPPT, dan Transtools untuk Bahasa Indonesia

Mesin Penerjemah …

- Dapat digunakan untuk menerjemahkan query atau dokumen
- *Performance* dari query yang diterjemahkan dengan mesin penerjemah berkisar antara 60-80% dibandingkan dengan *performance* dari monolingual

Keterbatasan dari Mesin Penerjemah

- Dasar dari mesin penerjemah adalah aturan linguistik sehingga hasilnya akan baik jika query ditulis dalam kalimat sesuai dengan tata bahasa yang baik.
- Biaya pembuatan mesin penerjemah sangat mahal.
- Seringkali tidak dapat menerjemahkan kata gabungan dan *proper nouns*.

Cara Penerjemahan

- Mesin Penerjemah
- Kamus Dwibahasa
- Korpus Paralel

Kamus Yang Dapat Dibaca oleh Mesin (*Machine Readable Dictionaries*)

- Berdasarkan pada kamus cetak dwibahasa
  - Tersedia secara luas
- Digunakan untuk menghasilkan daftar kata dwibahasa
  - Pemetaan kata pada Cross-language dapat dilakukan
    - Kadang diurutkan sesuai dengan penggunaannya yang paling umum
- Tantangannya adalah memilih terjemahan yang paling tepat

Penerjemahan Query

- Penerjemahan kata demi kata
  - Biasanya ada 5-10 terjemahan per kata
- Contoh kamus dwibahasa:
  - Collins
  - Kamus gratis : http://www.freedict.com
  - Babylon : http://www.babylon.com
  - Linguistic Data Consorsium
  - EuroWordNet

Masalah Penerjemahan dengan Kamus

- Hasilnya sekitar 50% dari monolingual (lebih jelek) karena satu kata dapat diterjemahkan dalam beberapa kata dalam bahasa lain.
- Masalah :
  - Arti kata sangat ambigu (ambiguity senses)
  - Penerjemahan frase
  - Cakupan kamus (kata tidak ada dalam kamus)
  - Singkatan
  - Kata gabungan

Contoh Penerjemahan Query

- Query Inggris nomor 22 (dari TREC):
  - The effects of chocolate on health?
- Versi Indonesia dari query di atas (Pengaruh permen coklat pada kesehatan) jika diterjemahkan ke Inggris mengguna-kan kamus Indonesia-Inggris:
  - Influence hard candy brown chocolate cocoa health

Contoh Penerjemahan Query

Indonesia		Inggris
pengaruh	→	influence
permen	→	hard candy, candy
coklat	→	brown, chocolate, cocoa
kesehatan	→	health

- Contoh lain : dari kamus Spanyol -Inggris
  - Chocolate (Spanyol)
  - chocolate-coloured; dark red; chocolate; drinking chocolate, cocoa, blood, dope, hash, pot (Inggris)

Contoh Penerjemahan Frase

- Query Inggris 1 (dari TREC):
  - Reasons for controversy surrounding Waldheim's World War II action
- Penerjemahan query dalam versi Indonesia dengan kamus Indonesia-Inggris:
  - Controversy measure action step waldheim world kingdom war battle II
- Frase yang tidak diterjemahkan akan kehilangan makna sesungguhnya (*perang dunia*)

Contoh Penerjemahan Frase

Indonesia		Inggris
kontroversi	→	controversy
tindakan	→	measure, step, action
waldheim	→	waldheim
perang	→	war, battle
dunia	→	world, kingdom
ii	→	ii

Masalah Penerjemahan Frase

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

Masalah Penerjemahan Frase

- Frase yang merupakan gabungan kata pada satu bahasa dapat diterjemahkan menjadi frase satu kata pada bahasa yang lain.
- Contoh:
  - South Africa (Inggris)
  - Südafrica (Jerman)

Hasil Penelitian

- Hasil dari query Inggris yang diterjemahkan ke Indonesia adalah 33% lebih buruk dari monolingual.
- Hasil dari query bahasa Indonesia yang diterjemahkan ke Inggris adalah 62% lebih buruk dari monolingual.

Mengatasi *Ambiguity*

- 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 dipilih sebagai kata terjemahan.
- Contoh:
  - Peudo relevance feedback
  - Local context analysis
  - Term Similarity
  - Probabilitas terjemahan kata

Mengatasi *Ambiguity* dengan Menggunakan *Part-of-Speech Taggers*

- Batasi terjemahan dengan *part of speech*
  - Kata benda, kata kerja, kata sifat, …
  - *Taggers* yang efektif tersedia untuk bhs Inggris
- Hasilnya baik bila kalimat querynya lengkap
  - Query pendek tidak memberikan dasar yang cukup untuk melakukan *tagging*
  - Pencocokan yang dibatasi (jenis katanya) dapat menurunkan monolingual IR
    - mis., kata benda pada query dapat merupakan kata kerja pada dokumen

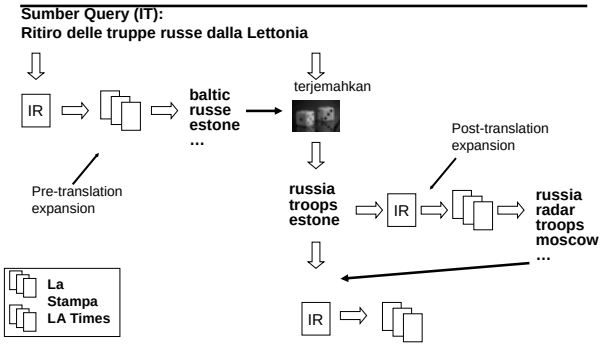
Mengindeks Frase

- Ada tiga cara untuk mengidentifikasi frase
  - Semantik (mis., yang muncul di kamus)
  - Sintaktik (mis., diperoleh sebagai frase kata benda)
  - Co-occurrence (kata yang sering muncul bersama)
- Hasil dari frase semantik cukup baik

Query Expansion dengan Pseudo Relevance Feedback

- 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 penerjemahan
- Kombinasi Pre- & Post-Translation QE
  - Menambahkan kata-kata pada query sebelum dan sesudah diterjemahkan
- Hasil: Kombinasi pre- dan post-translation query expansion yang terbaik; query pendek yang paling besar mendapat efek query expansion

Query Expansion pada CLIR



CLIR

Teknik Berdasarkan *Korpus*
Untuk Pelacakan *Free Text*

Jenis Korpus Dwibahasa

- Korpus paralel: koleksi berisi dokumen yang sama dalam beberapa bahasa
  - mis. korpus UN dalam bahasa Perancis, Spanyol & dan Inggris
  - Pasangan dokumen
  - Pasangan kalimat
  - Pasangan kata
- *Comparable corpora* (korpus yang sebanding)
  - Koleksi berisi dokumen yang topik, waktunya dll. sama
  - mis. Kantor berita Swiss melaporkan dalam bahasa Jerman, Perancis, Italia
  - Pasangan koleksi
  - Pasangan dokumen

Penggunaan Korpus

- Penerjemahan Query menggunakan Korpus yang *comparable*
  - 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

Membuat Korpus Paralel

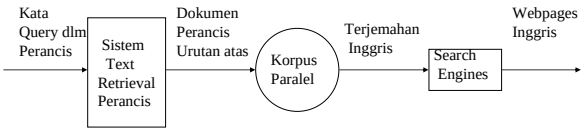
- Korpus paralel biasanya mempunyai domain yang sama
  - Mencari domain yang tepat sangatlah sukar
  - Mis. Dokumen PBB tersedia dalam beberapa bahasa, tetapi topiknya khusus dan jumlahnya terbatas.
- Alternatifnya : membuat sendiri
  - Mulai dengan korpus monolingual
  - Gunakan mesin penerjemah otomatis untuk bahasa kedua
  - Jika kesalahan menerjemahkan tidak mempengaruhi teknik IR

Membuat Korpus Paralel …

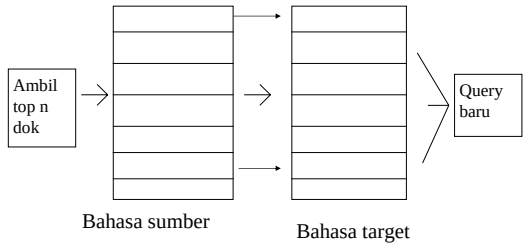
- Alternatif lain:
  - Mencari pasangan webpages di internet
  - Jika jumlahnya memadai, korpus dapat menghasilkan terjemahan yang baik.

Pseudo-Translation dengan *Parallel / Comparable Corpus*

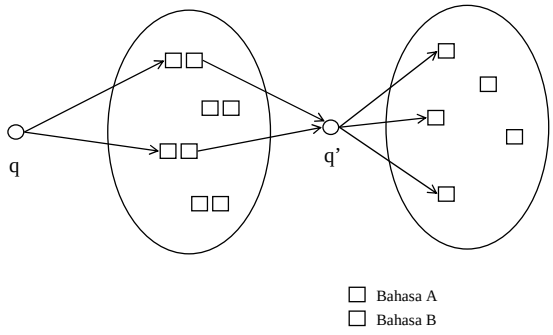
- Masukkan kata query dalam bahasa Perancis
- Ambil top dokumen Perancis pada korpus paralel
- Buat query dari terjemahannya dalam bhs Inggris
- Lakukan pelacakan *free text* monolingual



Pseudo-Translation ...



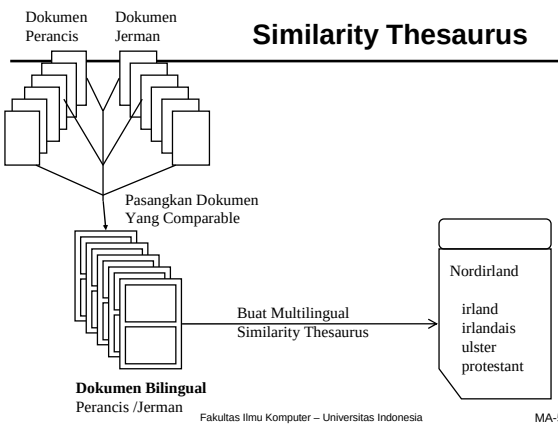
Pseudo-Translation



Similarity Thesaurus Dari Korpus Paralel

- *Similarity thesauri*
  - Ambil kata-kata yang dari segi terjemahannya sama dari korpus multilingual yang telah dipasangkan (pasangan dokumen)
  - Untuk setiap kata, cari yang paling mirip dengan bahasa lain
    - Gunakan hanya beberapa yang paling atas (5 atau lebih)
  - Catat kesamaannya pada thesaurus
  - Bergantung pada korpus yang kualitasnya baik

Similarity Thesaurus



Belajar dari Pasangan Dokumen

- Hitung seberapa seringnya setiap kata muncul sebagai pasangan
  - Anggap setiap pasangan sebagai satu dokumen

	Kata Inggris					Kata Indonesia			
	E1	E2	E3	E4	E5	ID1	ID2	ID3	ID4
Dok 1	4		2			2			1
Dok 2	8		4			4			2
Dok 3		2		2			2	1	
Dok 4		2	1				2		1
Dok 5	4				1	2		1	

Similarity Antara Dokumen dan Kata

- Dok 1 dan Dok 2 menggunakan pola kata yang mirip
  - Setelah menghitung jumlah kata di tiap dok
  - Cara kerja *vector space retrieval*
    - Hitung bobot dari tiap kata
    - *Cosine*: Normalisasi panjang, hitung *inner product*
- Kata pada E1 dan E3 digunakan dengan cara yang sama
  - Kata E1 & ID1 (atau E3 & ID4) lebih mirip
    - Penghitungan yang sama menghasilkan "*term similarity*"
    - Dapat dilakukan antar bahasa dan dalam bahasa itu sendiri

Korpus yang Dipasangkan Kalimatnya

- 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 penerjemahan query
  - Hasilnya baik bila domainnya sama
- Belum secara langsung digunakan untuk retrieval yang efektif
  - Tetapi semua eksperimen telah menyertakan *domain shift*

Penggunaan Korpus

- Masalah
  - Korpus yang sesuai sukar diperoleh
  - Korpus untuk latihan (*training*) harus sangat besar
  - Korpus cenderung bergantung pada aplikasi atau domain

Menggunakan Korpus yang Tidak Dipasangkan (*Unaligned Corpora*)

- Dokumen mengenai subyek dari set yang sama
  - Hubungan antara pasangan dokumen tidak diketahui
  - Mudah diperoleh pada banyak aplikasi
- Dua pendekatan
  - Gunakan kamus untuk terjemahannya
    - Perbaiki dengan korpus dwibahasa yang tidak dipasangkan
  - Gunakan kamus untuk menemukan pasangannya pada korpus
    - Lalu ambil kata terjemahannya dari pasangan itu

Mana yang Digunakan?

- *Controlled vocabulary*
  - Matang, efisien, mudah dijelaskan
- Berdasarkan kamus
  - Mudah, cakupannya luas
- Korpus yang *comparable* dan *paralel*
  - Efektif pada domain yang sama
- Korpus yang tidak dipasangkan
  - Eksperimen

Hasil CLIR

- Mesin Penerjemah
  - ~keefektifannya 80% dibandingkan monolingual pada domain umum
- Teknik Penggunaan Kamus
  - ~ keefektifannya 80% dibandingkan monolingual pada domain umum
- Teknik Penggunaan Korpus Paralel dan *Comparable*
  - ~ keefektifannya 80% dibandingkan monolingual pada domain umum
  - ~ keefektifannya 90% dibandingkan monolingual pada domain khusus

Kesulitan Utama pada CLIR

- Resources
- Sistem CLIR memerlukan *resources* yang dikembangkan dengan baik
  - Tools untuk Memroses Bahasa
  - *Resources* bahasa
- *Resources* sangat mahal untuk membuat, memelihara, dan memperbaiki

Apa itu CLEF?

- Cross-Language Evaluation Forum (CLEF) merupakan suatu kegiatan evaluasi CLIR untuk bahasa-bahasa Eropa
- Untuk meningkatkan partisipasi pada penelitian CLIR di Eropa

CLEF Tasks

- Multilingual
- Bilingual
- Monolingual (non Inggris)
- Koleksi dengan domain khusus

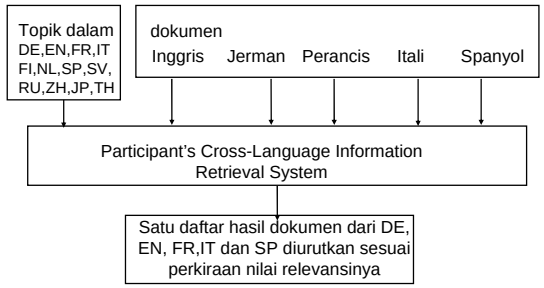
CLEF 2001 - Data Koleksi

- Multilingual comparable corpus yang terdiri dari dokumen surat kabar dan kantor berita nasional dalam 12 bahasa (DE,EN,FR,IT,NL,SP dll.). Lebih dari 1 juta dokumen.
- Query teridiri dari 50 topik dalam 12 bahasa Eropa (DE,EN,FR,IT,NL,SP+FI,RU,SV) and beberapa bahasa Asia (JP,CH, ID dll.)

Penelitian Multilingual Task

- Koleksi : dari surat kabar Inggris, Itali, Jerman dan Perancis
- Query : Inggris
- Hasil : suatu daftar dokumen yang relevan dari semua koleksi

CLEF – Multilingual IR



Tahapan MLIR

- Terjemahkan query menggunakan *resources* yang ada (mis. kamus gratis di internet) sesudah menghilangkan stopwords dan menerapkan stemmer.
- Lakukan teknik untuk menghilangan ambigu pada kata
- *Expand* query

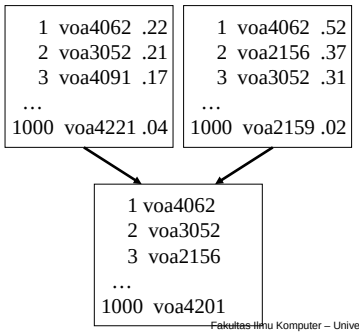
Tahapan MLIR …

- Proses query yang sudah diterjemahkan di tiap koleksi untuk semua bahasa
- Gabungkan hasilnya menjadi suatu daftar
- Masalah : menemukan urutan dokumen yang paling tepat untuk berbagai bahasa.

Pendekatan MLIR

- Pengindeksan
  - Satu indeks untuk tiap bahasa
    - Ada masalah penggabungan daftar dok (*merging*)
  - Satu indeks untuk semua bahasa
    - Tidak ada masalah dalam pembuatan list dok

Menggabungkan Daftar Urutan



- Penggabungan Daftar Urutan
  - Urutan (Rank)
  - Normalisasi Nilai dok
  - *Round robin*

Seleksi pada Multilingual

Query in English: `Swiss bank` Search

German Query: `(Swiss) (Bankgebäude, bankverbindung, bank)`

1 (0.91) U.S. Senator Warpathing	NZZ	June 14, 1997
2 (0.57) [Bankensecret] Law Change	SDA	August 22, 1997
3 (0.52) Swiss Bankers Criticized	AP	June 14, 1997
4 (0.36) Banks Pressure Existent	NZZ	May 3, 1997
5 (0.28) Bank Director Resigns	AP	July 24, 1997

Seleksi yang Sesuai dengan Bahasanya

Query in English: `Swiss bank` Search

English	German
	`(Swiss) (Bankgebäude, bankverbindung, bank)`
1 (0.72) Swiss Bankers Criticized AP / June 14, 1997	1 (0.91) U.S. Senator Warpathing NZZ / June 14, 1997
2 (0.48) Bank Director Resigns AP / July 24, 1997	2 (0.57) [Bankensecret] Law Change SDA / August 22, 1997
	3 (0.36) Banks Pressure Existent NZZ / May 3, 1997

Program Evaluasi CLIR

- **TIDES:** TDT (Topic Detection and Tracking) - Cina-Inggris tracks; Inggris/Perancis – Arabic
- **NTCIR:** Nat.Inst. for Informatics, Tokyo. Cina-Inggris; Jepang-Inggris CLIR tracks
- **AMARYLLIS:** CLIR untuk bahasa Perancis
- **CLEF:** evaluasi CLIR untuk bahasa Eropa