ELN 2006

Cross Language Information Retrieval

Metodi per il recupero di informazioni multilingua

Matteo Tanca

Laurea specialistica in Informatica – Università di Pisa

[tanca@cli.di.unipi.it]

Introduzione

- Il Cross Language Information Retrieval (CLIR) è una sottodisciplina dell' Information Retrieval
  - Fornisce metodi per il recupero di informazioni espresse in un insieme di linguaggi, a partire da una richiesta espressa in uno solo di essi
  - Si suddivide a sua volta in diversi campi:
    - Cross language text retrieval
    - Cross language speech retrieval
    - Cross language image retrieval
    - Interactive CLIR
    - Cross Language question answering
    - …

Sommario

1. CLIR: Descrizione Generale
    - Definizione, motivazioni, ambiti di ricerca
2. Text Retrieval
3. Image Retrieval
4. Question Answering
5. Interactive CLIR
6. Risorse utili
7. Bibliografia

Descrizione generale (1/5)

- Problema
  - Input
    1. Una collezione di informazioni, formalizzate in un determinato insieme di lingue
    2. Una richiesta, espressa in uno dei linguaggi
  - Output
    1. L'insieme di tutte le informazioni rilevanti (ai fini della richiesta) che si trovano nella collezione, a prescindere dalla lingua in cui sono formalizzate (tali informazioni non saranno necessariamente tradotte nella lingua adottata per la richiesta)

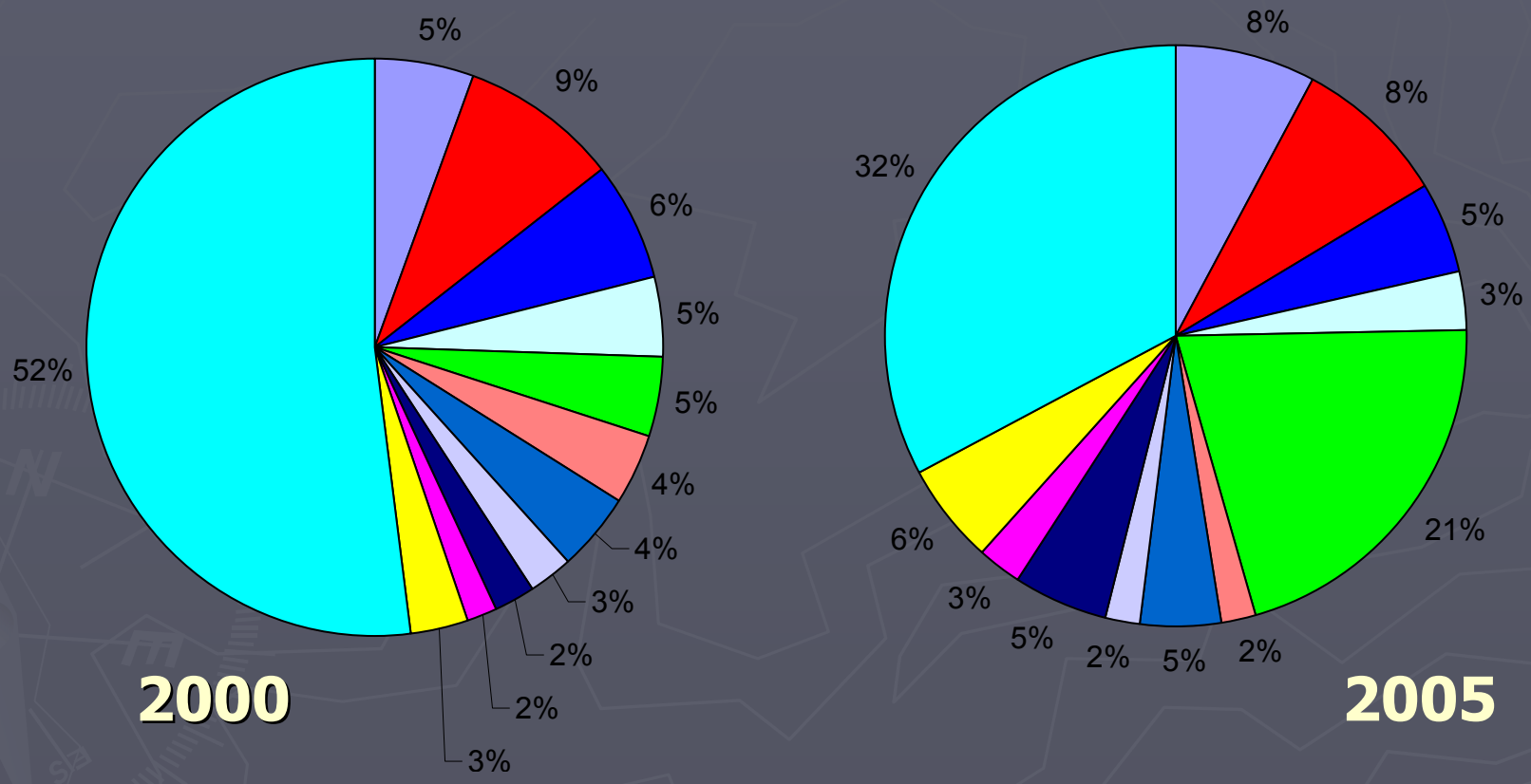
N.B.: si parla di informazioni, piuttosto che di documenti, in quanto a seconda del campo del CLIR considerato, si tratterà di documenti testuali, risposte a quesiti di tipo QA, trascrizioni di parlato…

Descrizione generale (2/5)

- Motivazioni
  - Grazie alla rapida diffusione di Internet, negli ultimi anni è cresciuta enormemente la disponibilità di informazioni, ormai non più disponibili soltanto in inglese…
  - L' inglese sta via via perdendo il tradizionale ruolo di "lingua del Web", perciò una buona padronanza dell'inglese non è più sufficiente ai fini di una ricerca completa
  - Un generico utente, affidandosi ai tradizionali metodi di ricerca monolingua ed all'uso della sola lingua madre, non ha la possibilità di usufruire della maggior parte delle informazioni disponibili (quantomeno nella maggioranza dei casi)

Descrizione generale (3/5)

- Distribuzione linguistica dell'utenza Internet



Fonte: Global Reach

Descrizione generale (4/5)

- Ambiti di ricerca del CLIR
  - Text retrieval: è stato il primo campo di ricerca, ed è perciò il più avanzato e documentato campo d'applicazione
  - Speech retrieval
  - Image retrieval: recupero di documenti in una collezione di immagini corredate di annotazioni (espresse in una o più lingue appartenenti ad un insieme di partenza). Data una immagine in input, in base alle meta-informazioni associate reperisce ogni altra immagine che possa essere considerata rilevante. Si tratta di un'applicazione particolarmente interessante in campo medico

Descrizione generale (5/5)

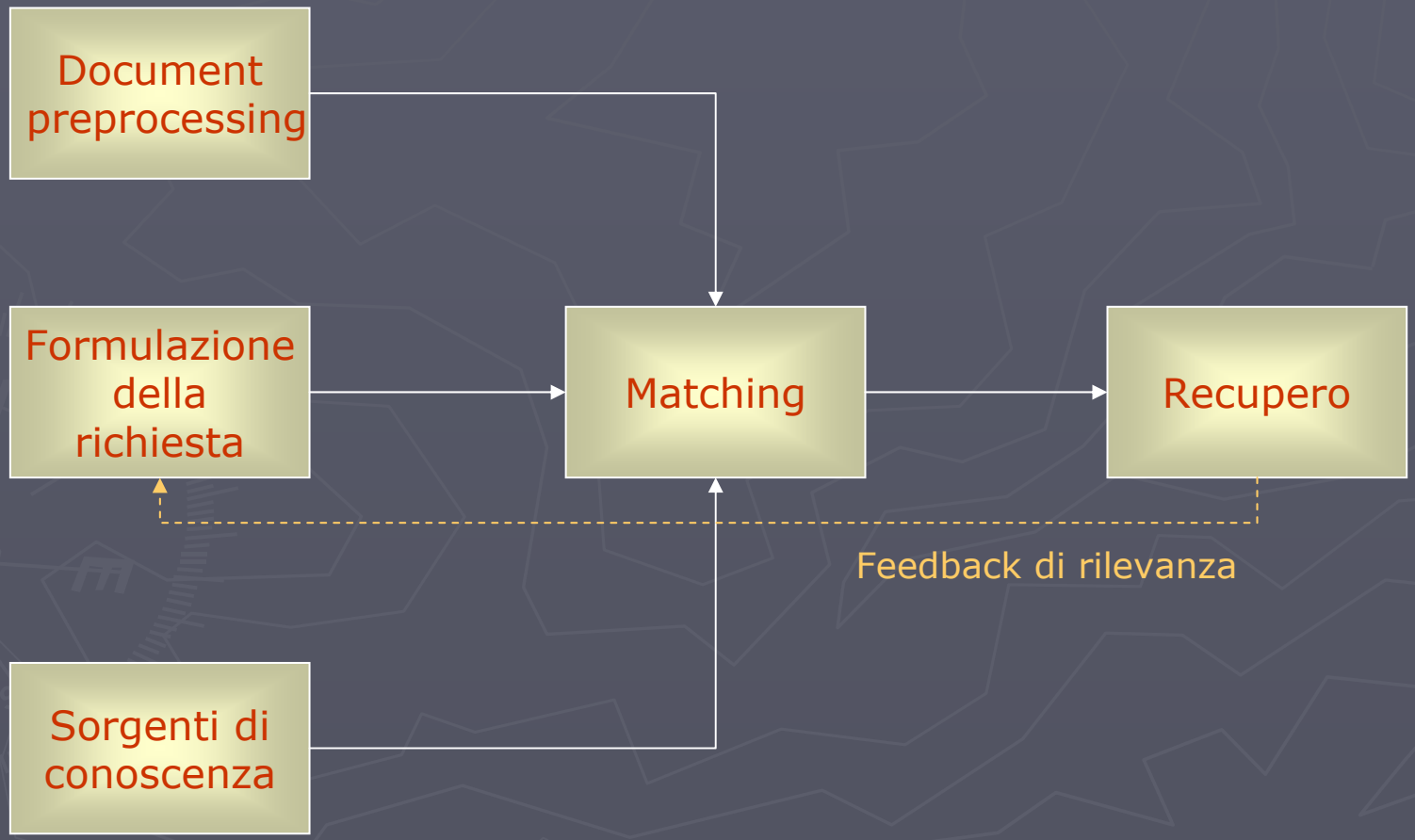
- Ambiti di ricerca del CLIR

  - Interactive CLIR: un approccio di ricerca che considera le applicazioni del CLIR da una prospettiva di tipo 'user inclusive', cercando di "*determinare come meglio assistere l'utente nella ricerca di informazioni espresse in linguaggi sconosciuti, piuttosto che determinare il miglior modo per un algoritmo di trovare informazioni espresse in linguaggi diversi da quello iniziale*" (iCLEF 2006 Guidelines)

  - Cross Language Question Answering: ricerca di risposte ad una domanda, posta in una singola lingua, all'interno di sorgenti di informazione multilingua. Tipicamente prevede la formulazione della risposta nella lingua usata per la domanda

Text retrieval (1/20)

- Modello concettuale del CL Text Retrieval



Text retrieval (2/20)

- Document preprocessing
  1. Identificazione lingua e uniformazione delle codifiche di carattere usate
     - Documenti scritti in lingue differenti possono essere codificati diversamente (GB, Big5, Latin1…)
     - Soluzione: codifica standard UNICODE (UTF-8, UTF-16)
  2. Tokenization
     - Stemming, separazione di parole composite, riconoscimento di sintagmi nominali etc.
  3. Rimozione stopwords e normalizzazione
     - Stopwords: parole comuni, che non contribuiscono alla rilevanza di un documento e non sono considerate nelle richieste (ad esempio 'Io', 'a', 'un', 'o' etc. etc.)

Text retrieval (3/20)

- Formulazione della richiesta
  - Vocabolari controllati e thesauri multilingua (prodotti manualmente)
    - La formulazione della richiesta risulta complessa per utenti inesperti
    - Problemi di manutenzione e aggiornamento
  - Formulazione libera
    - Due tipologie: con traduzione della richiesta o senza
      Con traduzione della richiesta: problemi di ambiguità (che cresce all'aumentare del numero di lingue considerate contemporaneamente)
    - Senza traduzione: approccio numerico automatico LSI

Text retrieval (4/20)

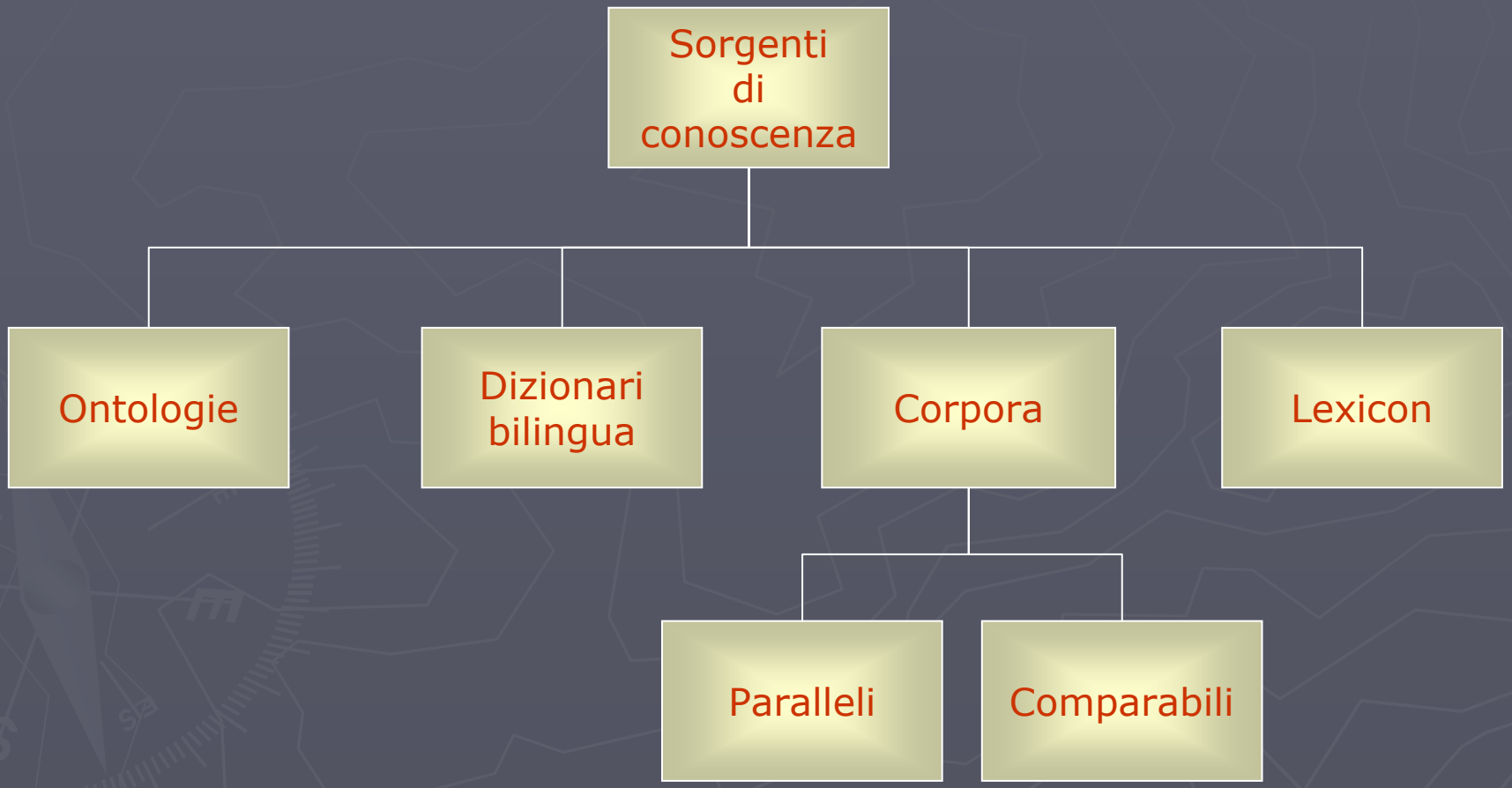
- Formulazione della richiesta
  - Formulazione libera
    - Traducendo la richiesta si va incontro ad ambiguità ed errori di traduzione
    - Il problema può essere risolto, o almeno limitato, mediante ritraduzione della richiesta o scelta interattiva del termine tradotto
    - La prima soluzione permette ad un utente che ignori la lingua in cui è stata tradotta la richiesta di eliminare quei casi in cui la scelta dei termini, in base alla loro ritraduzione, appaia errata

Text retrieval (5/20)

- Formulazione della richiesta
  - Formulazione libera
    - La seconda soluzione, valida soltanto per utenti che abbiano una conoscenza almeno approssimativa della lingua in cui è stata tradotta la richiesta, permette di evitare casi di ambiguità nella traduzione
    - L'approccio senza traduzione della richiesta è invece utilizzato da tecniche interlinguistiche, ad esempio il Latent Semantic Indexing, nel quale viene generata una rappresentazione della richiesta indipendente dal linguaggio di formulazione

Text retrieval (5/20)

- Sorgenti di conoscenza

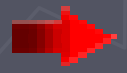


Text retrieval (6/20)

- Sorgenti di conoscenza
  - Ontologie
    - Codificano relazioni tra concetti
    - I thesauri sono ontologie finalizzate all'utilizzo nell'IR, che codificano un determinato dominio di conoscenza
    - I thesauri multilingua rappresentano una sorgente di conoscenza fondamentale nei sistemi di CLIR
  - Dizionari bilingua automatici
    - Vengono usati nella traduzione delle richieste. Ad un termine è associata una lista di traduzioni possibili, eventualmente corredate da informazioni di contesto
    - La scelta della traduzione può essere arbitraria (approccio stocastico), essere guidata dal contesto scelto per altri termini o da un eventuale ordine preferenziale all'interno della lista

Text retrieval (7/20)

- Sorgenti di conoscenza
  - Lexicon
    - Codificano la conoscenza necessaria ad un sistema di traduzione automatica ai fini dell'analisi automatica, della traduzione e generazione di linguaggio naturale
    - Sono progettati per risolvere in maniera automatica i problemi di ambiguità
    - L'uso di sistemi di traduzione automatica nel campo del CLIR è efficiente nel caso si traducano interi documenti, mentre è tipicamente controproducente nel caso si traducano le richieste, poiché la scelta di una singola traduzione da parte del sistema produce generalmente risultati abbastanza scadenti nella ricerca

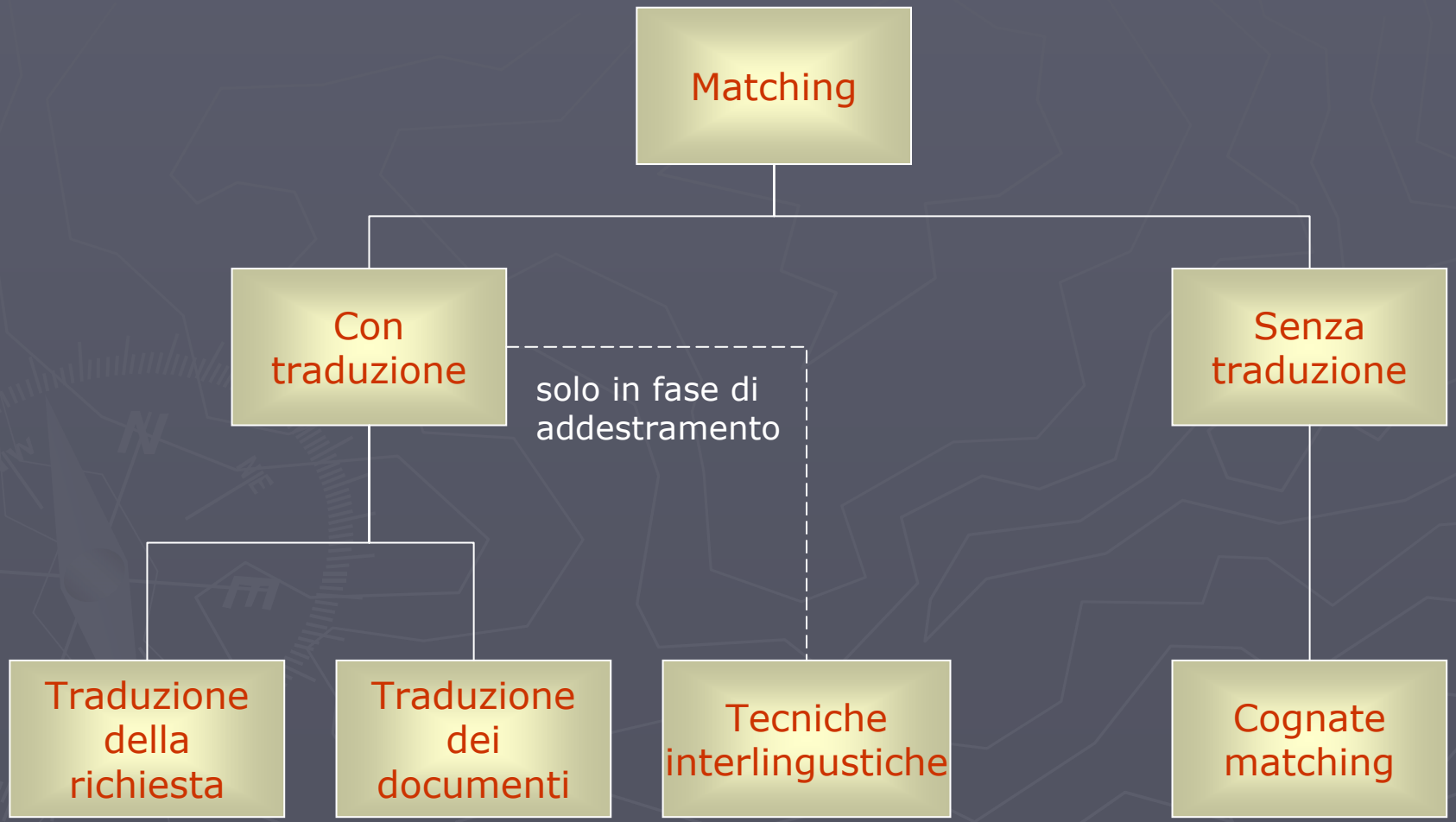


Text retrieval (8/20)

- Sorgenti di conoscenza
  - Corpora
    - Collezioni di (estratti di) testi in una o più lingue, eventualmente annotati
    - Paralleli: collezioni di testi, ognuno presente in uno o più lingue. La lingua originale del documento non deve necessariamente essere la stessa per ognuno di essi
    - Comparabili: insiemi di documenti simili in più di una lingua. Il concetto di somiglianza si riferisce all'ambito dei testi

Text retrieval (9/20)

- Matching



Text retrieval (10/20)

- Matching
  - Senza traduzione: cognate matching
    - Ricerca corrispondenze fra i termini in base a similitudini sintattiche o di pronuncia
    - Applicabile soltanto a lingue che abbiano radici comuni (un approccio simile non avrebbe nessun senso se fosse ad esempio applicato ad inglese e cinese…)
    - Benché utilizzabile anche da solo, questo approccio è tipicamento inserito in metodologie ibride, che applicano anche metodi di traduzione
  - Con traduzione: traduzione della richiesta
    - Poco costosa dal punto di vista computazionale
    - Problemi di forte ambiguità (il contesto per poter disambiguare è quasi nullo)
    - Efficace qualora sia possibile stabilire con esattezza il contesto dei termini (e di conseguenza la traduzione)

Text retrieval (11/20)

- Matching
  - Con traduzione: traduzione dei documenti
    - Molto costosa in termini di risorse computazionali (eventualmente impraticabile se il numero di documenti considerati e/o il numero di lingue previste è molto elevato)
    - Meno soggetta ad ambiguità della traduzione della richiesta (il contesto è tipicamente sufficientemente ampio per disambiguare)
    - Utile, quando computazionalmente praticabile, se si vogliono fornire direttamente documenti tradotti
  - Con traduzione in fase iniziale: tecniche interlinguistiche
    - Documenti e richieste sono rappresentati in maniera indipendente dal linguaggio
    - Latent Semantic Indexing (LSI): utilizzato negli approcci automatici

Text retrieval (12/20)

- Latent Semantic Indexing
  - Nota anche come Latent Semantic Analysis, è una tecnica per la costruzione automatizzata di uno spazio semantico
    - Lo spazio semantico dei termini è rappresentato tramite matrici dette termine-documento, che esprimono una relazione di occorrenza fra termini ritenuti significativi e documenti analizzati
  - Applicata nel campo del Cross Language Text Retrieval
    - Non necessita di traduzioni automatiche per le ricerche
    - Necessita di una fase di addestramento mediante corpora paralleli per la costituzione dello spazio semantico multilingua

Text retrieval (13/20)

- LSI: fasi

Text retrieval (14/20)

- Matrice termine-documento
  - Esprime il numero di occorrenze di termini in una collezione di documenti
    - Le colonne corrispondono ai documenti presi in esame
    - Le righe corrispondono ai termini considerati
    - Data una matrice termine-documento A, in posizione A[i, j] si avrà il numero di occorrenze dell'i-esimo termine nel j-esimo documento
    - Per dare maggior rilievo ad alcuni termini, la funzione di valutazione del numero di occorrenze è tipicamente pesata piuttosto che lineare
    - L'attinenza di un documento ad un certo argomento è definita in base al numero di occorrenze di alcuni termini

Text retrieval (15/20)

- Fattorizzazione SVD
  - Permette di decomporre una matrice rettangolare nel prodotto di tre matrici
    - Passaggio algebrico necessario per l'applicazione della riduzione del rumore presente nella matrice originale
    - La matrice A di partenza è fattorizzata in $A = TSD^T$
    - T e D sono le matrici ortogonali dei vettori singolari sinistri e destri
    - S è la matrice diagonale dei valori singolari
  - Un teorema garantisce l'esistenza della fattorizzazione SVD per una generica matrice

Text retrieval (16/20)

- Eliminazione del rumore
  - Genera un' approssimazione della matrice iniziale
    - Necessaria per rendere la matrice trattabile in termini computazionali
    - Riduce la dimensione della matrice, scartando i valori singolari più bassi nella matrice S della decomposizione e mantenendo solo i corrispondenti vettori delle matrici ortogonali T e D
    - Da un punto di vista semantico l'operazione può essere vista come una riduzione del "rumore di fondo" presente nella matrice iniziale, in quanto tende ad accorpare le dimensioni relative a sinonimi e a separare in dimensioni differenti le diverse accezioni dei termini polisemi

Text retrieval (17/20)

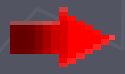
- Applicazione del LSI al CL Text Retrieval
  - Fase di addestramento
    - Necessaria per rendere la matrice trattabile in termini computazionali
    - Si sceglie un insieme iniziale di documenti, che vengono tradotti (in maniera automatica o manualmente per una traduzione più accurata), al fine di ottenere un corpus parallelo
    - L'analisi LSI del corpus produce uno spazio semantico multilingua
    - Nello spazio semantico prodotto, i termini che denotano esattamente la stessa entità (ad esempio nomi di Stati, fiumi etc.) avranno identica rappresentazione, mentre termini frequentemente associati per traduzione (ad esempio 'non' e 'not') avranno rappresentazioni simili

Text retrieval (18/20)

- Applicazione del LSI al CL Text Retrieval
  - Fase di popolamento dello spazio semantico
    - Lo spazio, in seguito alla fase di addestramento, può essere popolato mediante documenti monolingua, senza effettuare traduzione, sulla base dei loro termini più significativi
    - La locazione spaziale dei documenti inseriti sarà tanto più prossima a quella di altri documenti quanto più tali documenti saranno simili ad essi
    - La misura tipicamente adottata per verificare l'attinenza tra due documenti è la distanza coseno fra di essi, che risulta inversamente proporzionale alla loro affinità

Text retrieval (19/20)

- Applicazione del LSI al CL Text Retrieval
  - Formulazione della richiesta
    - La richiesta è formulata in linguaggio naturale, in uno qualsiasi dei linguaggi previsti dal sistema
    - La scelta del linguaggio in cui formulare la richiesta è del tutto arbitraria, in quanto la sua rappresentazione geometrica sarà indipendente dal linguaggio
    - La rappresentazione geometrica della richiesta, anche detta pseudo-documento, viene scalata sulla base dei valori singolari salvati durante la fattorizzazione SVD
  - Recupero dei documenti
    - In base alla distanza coseno fra la proiezione dello pseudo-documento ed i documenti più prossimi ad essa, si genera una lista dei documenti ritenuti attinenti, ordinata per rilevanza in base alla distanza



Text retrieval (20/20)

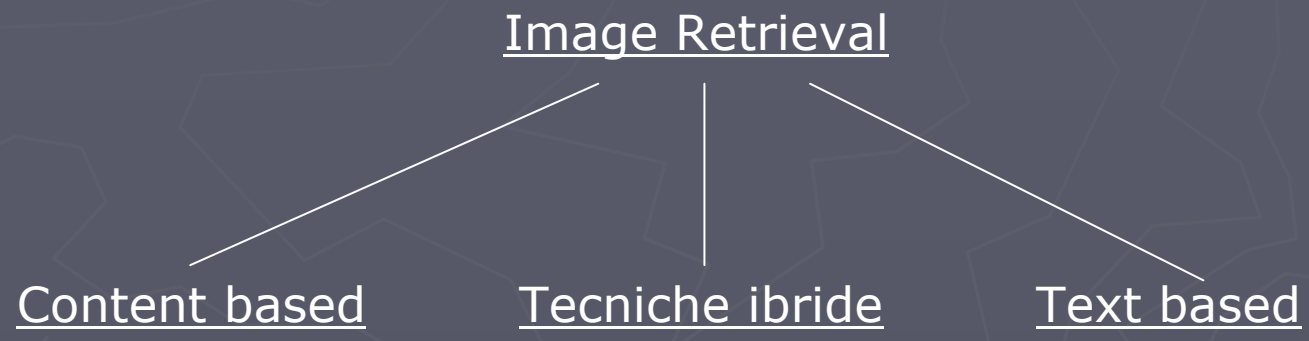
- Applicazione del LSI al CL Text Retrieval
  - Feedback di rilevanza
    - Metodo iterativo per l'espansione della richiesta iniziale
    - In base ai termini contenuti nei documenti più rilevanti trovati dopo la prima ricerca, viene generato un nuovo pseudo-documento, utilizzato per ottenere una lista più ampia ed in generale migliore della precedente, in termini di precisione e richiamo
    - Il procedimento viene ripetuto fino al raggiungimento di una determinata soglia (ad esempio il numero di documenti trovati al di sopra di una certa rilevanza)

Image retrieval (1/15)

- Motivazioni
  - Descrizione testuale delle immagini
    - Le immagini sono spesso accompagnate da didascalie e annotazioni, quali autore, titolo, parole chiave etc. etc.
    - E' possibile reperire immagini mediante una ricerca testuale
    - Le immagini sono rilevanti a prescindere dalla lingua in cui sono espresse le annotazioni, motivo in più per non limitare la ricerca ad una sola lingua…
  - Forte interesse sia scientifico sia mediatico
    - Evidente utilità nel campo medico (grandi quantità di materiale visivo, attualmente scarsamente fruibile)
    - Interesse da parte dell'utente comune (necessità di reperire informazioni non solo testuali…)

Image retrieval (2/15)

- Schema concettuale Image Retrieval



- Text based
  - Opera su collezioni di immagini opportunamente annotate
  - Il processo di annotazione è piuttosto costoso, in quanto necessita di intervento umano
  - La richiesta da parte dell'utente è testuale, quindi sono validi problemi e tecniche viste nel caso del CL text retrieval

Image retrieval (3/15)

- Content based
  - Opera su collezioni di immagini, senza il bisogno di alcuna annotazione
  - I sistemi più comuni fanno uso di caratteristiche a basso livello, come texture, forme, distribuzione del colore e simili
  - Alcuni sistemi specifici sono progettati per riconoscere forme peculiari di un determinato dominio
  - Sono possibili i seguenti tipi di richieste: richiesta mediante esempio (fornita dall'utente o scelta da un insieme casuale), richiesta mediante abbozzo (l'utente fornisce una approssimazione grafica sommaria dell'immagine che sta cercando ed il sistema determina le immagini simili in base ad un'analisi delle caratteristiche a basso livello)
  - Se previsto, l'utente può migliorare la ricerca mediante un feedback di rilevanza, marcando i risultati come rilevanti, irrilevanti o neutri ed iterando la ricerca

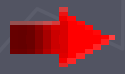


Image retrieval (4/15)

- Tecniche ibride
  - Cercano di combinare gli aspetti vantaggiosi delle due tecniche pure, limitandone i difetti

Text based	Content based
Vantaggi 1) Semplicità di costruzione della richiesta 2) Facilità di rappresentazione degli aspetti semantici Svantaggi 1) Possibili errori di traduzione 2) L'annotazione è soggetta ad errori, specie se realizzata automaticamente	Vantaggi 1) Non necessita di traduzione 2) Non richiede annotazione Svantaggi 1) La rappresentazione degli aspetti semantici non è intuitiva 2) E' necessario trovare o produrre delle immagini per le richieste

Image retrieval (5/15)

- Approcci ibridi
  - Approccio parallelo
    - Applica separatamente le due tecniche e opera una fusione dei risultati
  - Pipeline
    - Effettua una prima ricerca mediante tecnica text based o content based, filtrando i risultati applicando la tecnica rimanente
  - Transformation based
    - Estrae le relazioni fra immagini e testo, così da poterle utilizzare per convertire l'informazione testuale in indicazioni grafiche e viceversa

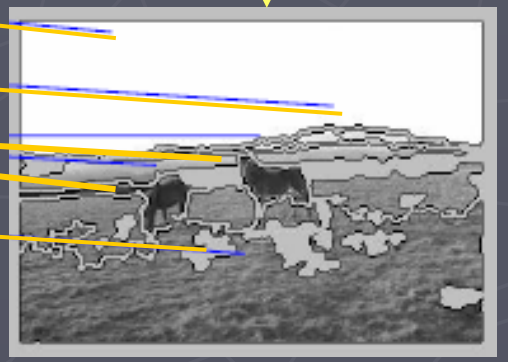
Image retrieval (6/15)

- Approccio transformation based (ImageCLEF 2004)
  - Le richieste testuali sono trasformate automaticamente nelle corrispondenti indicazioni grafiche per la ricerca content based, in due passi:
    1) Si estraggono le relazioni fra immagini e testi associati della collezione di ricerca
       - Divide le immagini in elementi grafici più piccoli
       - Associa i termini nell'annotazione ai corrispondenti elementi grafici
       - Si tratta di un procedimento analogo all'allineamento in un corpus parallelo, che produce una sorta di vocabolario grafico-testuale
    2) Si trasforma la richiesta testuale utilizzando il vocabolario generato al termine della prima fase

Image retrieval (7/15)

- Esempio di impostazione delle relazioni

suddivisione



Cielo → 01

Collina → 02

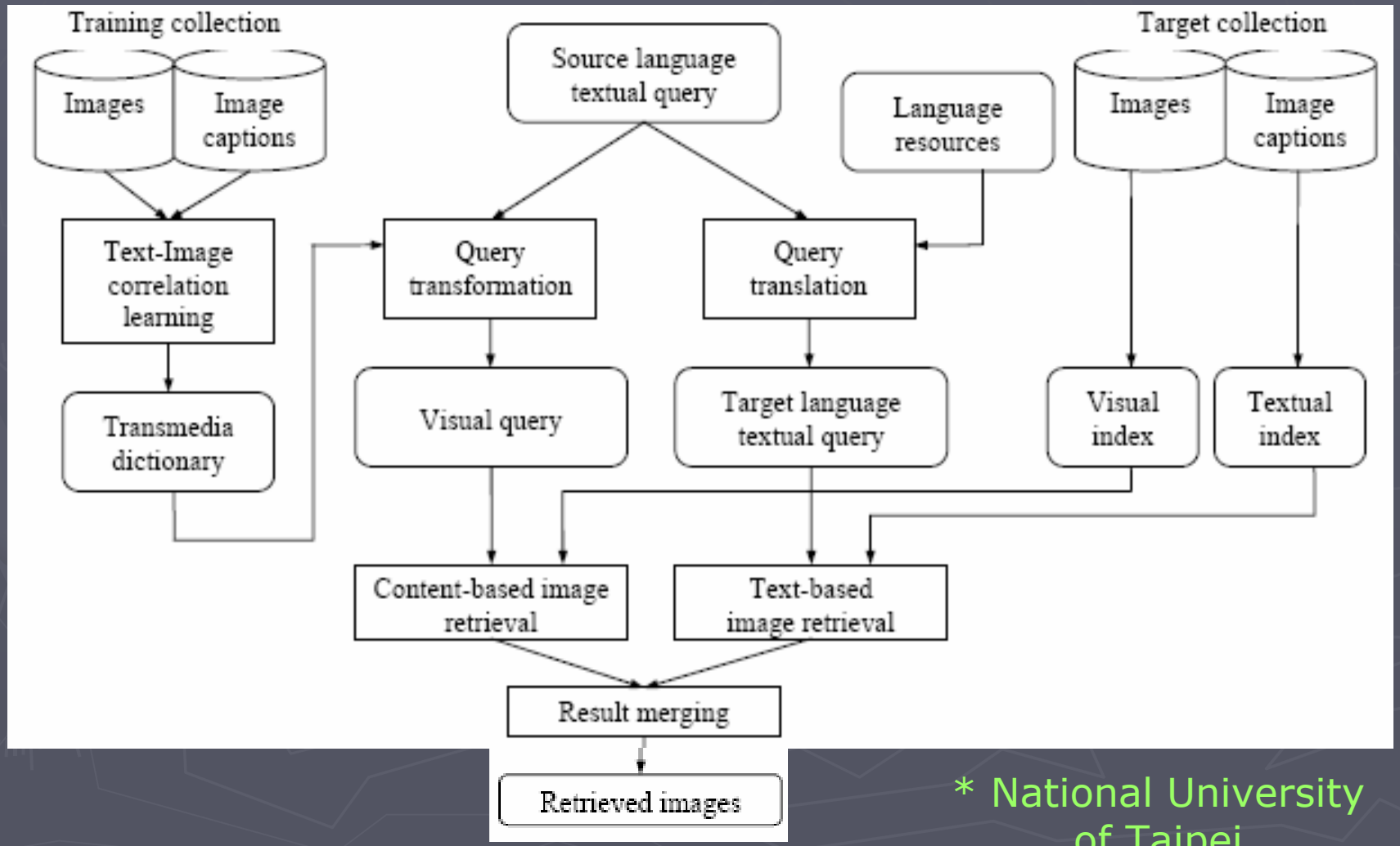
Puledro → 03

Cavallo → 03

Campo → 04

Image retrieval (8/15)

- NTU* @ ImageCLEF 2004: schema concettuale



* National University of Taipei

Image retrieval (9/15)

- NTU @ ImageCLEF 2005
  - Un evoluzione del precedente approccio
    - La richiesta iniziale è esclusivamente grafica
    - Il risultato della ricerca è composto di immagini e descrizioni, che vengono trattate come documenti allineati
    - Seleziona alcuni termini dalle annotazioni, così da generare un feedback di rilevanza testuale
    - La nuova richiesta, costruita in base al feedback, è di tipo testuale
    - Il prodotto finale è una ricerca ibrida, che combina l'approccio basato su trasformazione delle richieste con quello pipeline

Image retrieval (10/15)

- NTU @ ImageCLEF 2005: schema concettuale

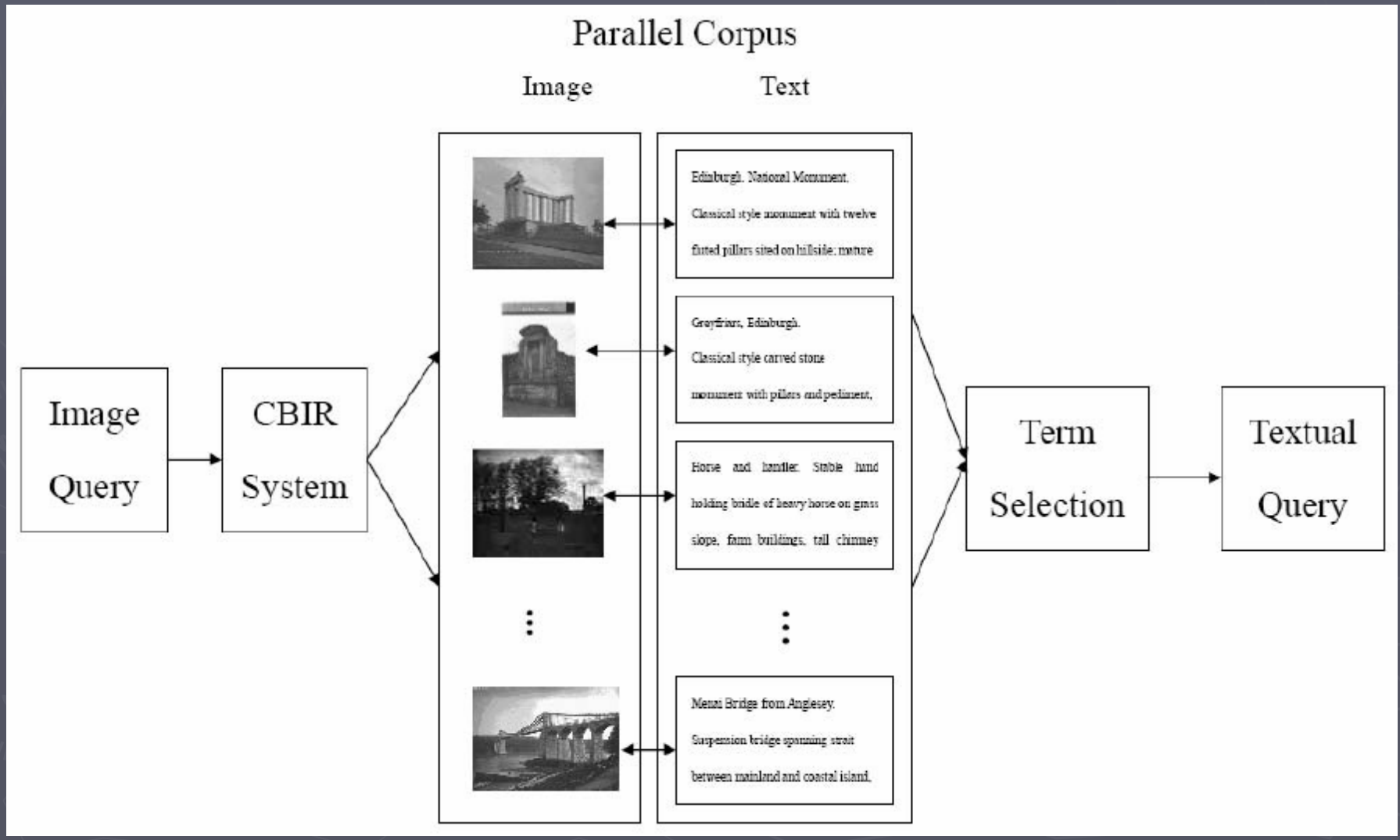


Image retrieval (11/15)

- NTU @ ImageCLEF 2005: dati e risultati
  - Dati
    - Una collezione di 28133 fotografie provenienti dalla biblioteca dell'università scozzese St. Andrews
    - Le immagini sono accompagnate da descrizioni testuali in lingua inglese
    - L'insieme di test contiene 28 argomenti, ciascuno rappresentato da una coppia <immagine, annotazione> (nel caso specifico l'annotazione è in lingua cinese)
  - Risultati
    - I risultati denotano una maggior precisione rispetto ad un utilizzo esclusivo di tecniche text o content based, in quanto si limita il problema delle traduzioni errate e della perdita di risultati a causa di sinonimia rispetto ai termini ricercati, mantenendo comunque i vantaggi di una ricerca basata su testo

Image retrieval (12/15)

- UB* @ ImageCLEF 2005: Medical Image Retrieval
  - Applicazione di tecniche ibride ad un contesto specifico
    - Combinazione di sistemi preesistenti: GIFT + SMART + MetaMap
    - GIFT (GNU Image Finding Tool): un sistema open source per la ricerca di immagini mediante tecniche content based
    - SMART: sistema per la ricerca in documenti testuali, sviluppato alla Cornell University
    - MetaMap: software sviluppato in ambito medico, che permette di associare testi a concetti biomedici, descritti dall'Unified Medical Language System (UMLS) Metathesaurus, una delle maggiori risorse semantiche elettroniche multilingua disponibili in campo medico

* State University of New York at Buffalo

Image retrieval (13/15)

- UB @ ImageCLEF 2005: metodologia
  - Tecnica ibrida coadiuvata da conoscenza specifica
    - Richieste iniziali costituite da una breve descrizione, accompagnata da una o più immagini d'esempio
    - Mediante MetaMap viene elaborata la parte testuale, per ottenere una lista di concetti UMLS
    - Mediante il Metathesaurus si reperiscono i termini che sono associati ai concetti nelle altre lingue di interesse
    - La lista di termini è fornita come richiesta per la ricerca mediante SMART
    - Le immagini iniziali sono usate come richiesta per GIFT
    - I 10 risultati migliori di entrambe le ricerche sono uniti ed utilizzati per espandere la richiesta iniziale
    - L'espansione è di tipo incrociato: i termini associati alle immagini fornite da GIFT sono usati per espandere la richiesta SMART e viceversa

Image retrieval (14/15)

- UB @ ImageCLEF 2005: schema concettuale

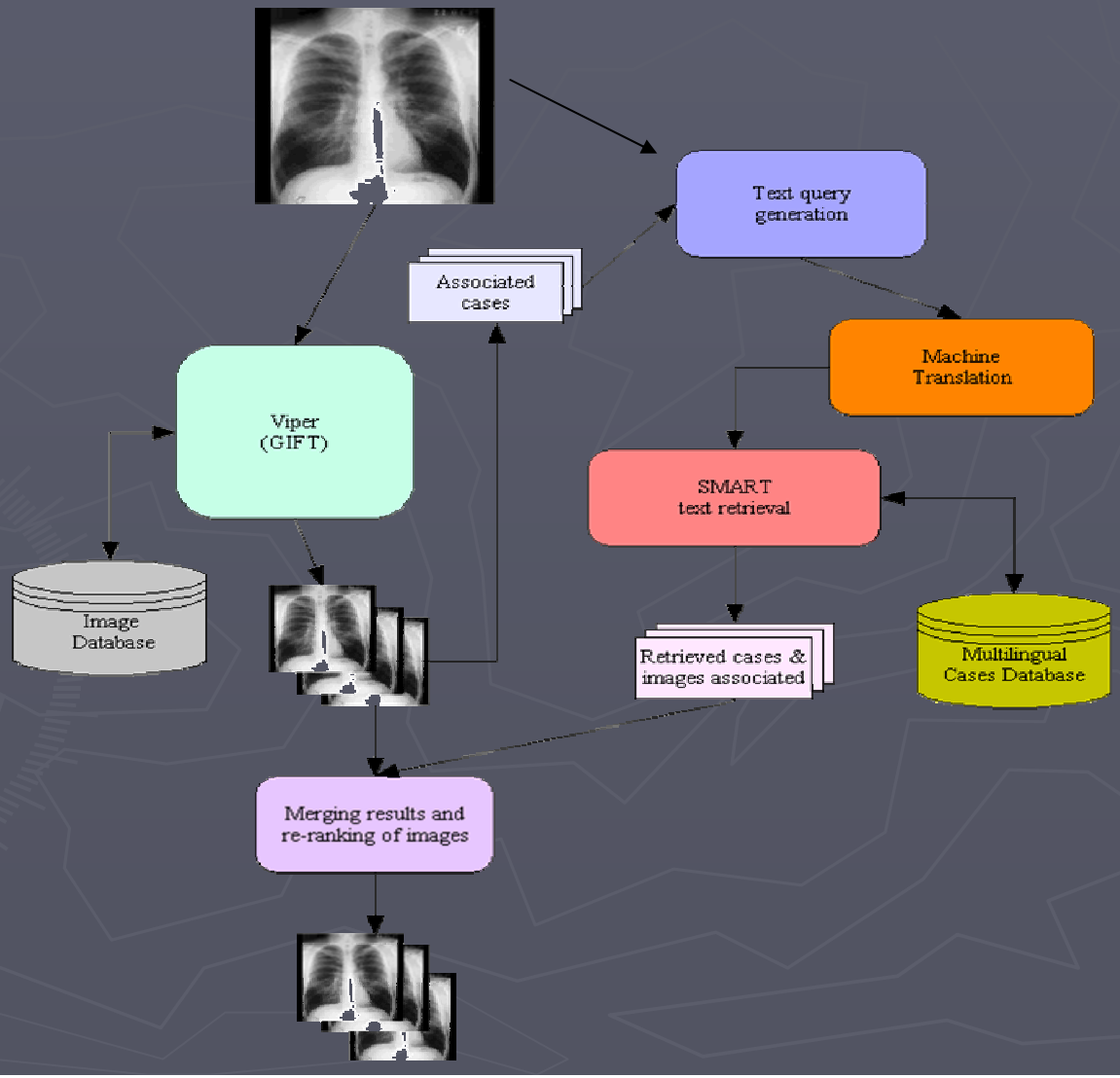


Image retrieval (15/15)

- UB @ ImageCLEF 2005: dati e risultati
  - Dati
    - Immagini proveniente da 4 differenti collezioni annotate (Casimage, MIR, Pathopic, PIER)
  - Risultati
    - 5 prove di ricerca, effettuate variando alcuni parametri, quali il peso dei risultati proveniente da SMART e GIFT, il numero massimo di termini da espandere durante la generazione delle richieste ed il numero di iterazioni da effettuare mediante feedback di rilevanza
    - Miglior rapporto trovato: rilevanza dei risultati di provenienza testuale "pesata" 3:1 rispetto ai risultati provenienti dalla ricerca content based ed espansione delle richieste coi 50 termini più affini
    - Risultati: +35% recall rispetto ad una ricerca solo testo, +150% rispetto ad una ricerca solamente visiva

Cross Language QA (1/19)

- CL Question Answering: definizione
  - Question Answering
    - Data una collezione di documenti, un sistema QA deve essere in grado di reperire risposte a domande poste in linguaggio naturale
    - E' un procedimento più complesso rispetto al semplice text retrieval, perché il risultato della ricerca dovrebbe essere non una lista di documenti rilevanti, ma una risposta precisa
    - Da molti considerati il futuro dei motori di ricerca, i sistemi QA devono confrontarsi con molte tipologie di domande (fattuali, definizioni, "Come…?", "Perché…?", ipotetiche etc. etc.)

Cross Language QA (2/19)

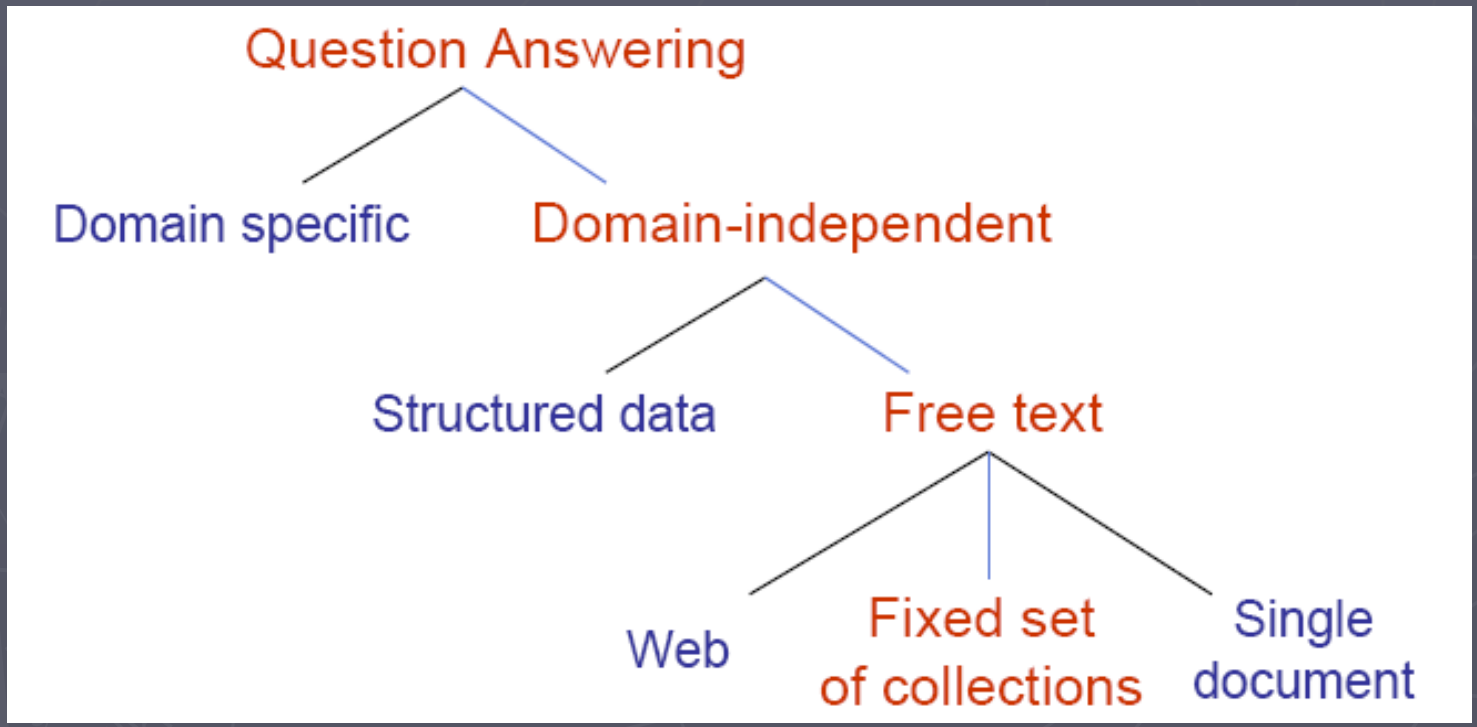
- CL Question Answering: definizione
  - Question Answering
    - Le sorgenti di informazioni possono essere di vario genere: collezioni di testi, archivi giornalistici, il Web…
  - Closed Domain Question Answering
    - Si occupa di domande riguardanti un particolare dominio di conoscenza
    - Può essere visto come una semplificazione del QA, in quanto il sistema può sfruttare conoscenze specifiche, tipicamente formalizzate in ontologie
  - Open Domain Question Answering
    - Si occupa di rispondere a domande generiche, senza alcuna restrizione del dominio di conoscenza
    - Può avvalersi di ontologie generiche (WordNet e simili)

Cross Language QA (3/19)

- CL Question Answering: definizione
  - CL Question Answering
    - Si tratta di una versione del Question Answering in cui la ricerca della risposta è estesa anche a documenti espressi in lingue differenti da quella in cui è stata espressa la domanda
    - La risposta trovata, se espressa in una lingua differente da quella della domanda, può essere fornita così com'è o essere tradotta automaticamente
    - I sistemi di CL QA sono spesso sistemi QA preesistenti opportunamente equipaggiati di meccanismi di traduzione, eventualmente modificati per tenere conto delle problematiche tipiche del CLIR

Cross Language QA (4/19)

- CLEF 2005 QA: schema concettuale del task



Cross Language QA (5/19)

- CLEF 2005 QA: caratteristiche
  - Sorgenti di informazioni
    - Corpora multilingua, costituiti da notizie del 1994-1995
    - Utilizzo opzionale del Web da parte di alcuni sistemi partecipanti, per cercare di verificare l'esattezza della risposta generata
  - Tipologie di domande proposte
    - 200 domande, di cui il 50% fattuali, il 25% definizioni, il 15% restrizioni temporali ed il 10% senza risposta
  - Risposte
    - Richieste risposte esatte, senza necessità di traduzione
    - Ogni risposta data è valutata come: corretta, del tutto sbagliata, inesatta o non supportata (qualora la risposta sia corretta ma non sia fornito l'identificativo della fonte)

Cross Language QA (6/19)

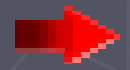
- CLEF 2005 QA: caratteristiche
  - Valutazione dei sistemi
    - Una misura adottata per la valutazione dei sistemi è l'accuratezza, data da #risposte esatte/totale domande
    - Un'altra misura è il punteggio pesato di confidenza, dato dalla formula

$$CWS = \frac{1}{Q} \sum_{i=1}^{Q} \frac{\text{number of correct responses in first i ranks}}{i}$$

dove Q è il totale delle domande

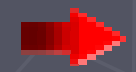
Cross Language QA (7/19)

- CLEF 2005 QA: approcci proposti
  - Uso di strumenti e risorse linguistiche
    - La maggioranza dei sistemi si è avvalsa di tecniche linguistiche quali: POS tagging, riconoscimento dei sintagmi nominali e simili
    - Parsing superficiale (shallow): alcuni sistemi hanno adottato tale forma di parsing, che consiste nell'utilizzo di parole chiave per la ricerca di passaggi interessanti all'interno dei documenti, per poi filtrare i risultati con espressioni regolari in base a criteri di somiglianza alla domanda (tale tecnica risulta efficiente per domande fattuali)



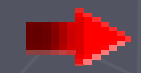
Cross Language QA (8/19)

- CLEF 2005 QA: approcci proposti
  - Uso di strumenti e risorse linguistiche
    - Parsing in profondità (deep parsing): molti sistemi hanno preferito una forma di parsing approfondito, che prevede un'analisi preventiva della collezione di documenti, per costruire alberi di dipendenza fra i documenti e generare liste dei modelli di risposte presenti
  - Approcci semantici
    - Diversi partecipanti hanno utilizzato tecniche per la associazione delle parole chiave presenti nella domanda a concetti semantici, tipicamente sfruttando WordNet o simili



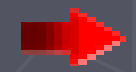
Cross Language QA (9/19)

- CLEF 2005 QA: approcci proposti
  - Approcci semantici
    - Qualche sistema ha utilizzato metodi di indicizzazione semantica (sul modello LSI), per poter applicare metodi numerici alla ricerca
  - Ricerca di fonti alternative
    - Alcuni partecipanti hanno previsto delle ricerche Web a tempo d'esecuzione, per cercare la risposta su portali specializzati (nel caso di identificazione di un particolare dominio di conoscenza nella domanda) o verificare la risposta generata



Cross Language QA (10/19)

- CLEF 2005 QA: approcci proposti
  - Traduzione
    - Praticamente tutti i sistemi hanno integrato per la parte di traduzione sistemi preesistenti, senza apportarvi modifiche sostanziali
    - Alcuni partecipanti hanno optato per la traduzione parola per parola della domanda, altri per la sola traduzione delle parole chiave
- CLEF 2005 QA: valutazione
  - Le risposte sono state valutate da giurie di esperti interne al CLEF, che hanno tenuto conto principalmente della correttezza e dell'esattezza delle risposte, come definite di seguito



Cross Language QA (11/19)

- CLEF 2005 QA: valutazione
  - Correttezza ed esattezza
    - Una risposta è corretta quando chiara nella forma e contenente l'informazione richiesta
    - L'esattezza riguarda invece la quantità di informazioni presenti nella risposta ed è inversamente proporzionale al numero di nozioni irrilevanti, rispetto alla domanda posta, contenute nella risposta
    - Data la domanda *Chi è il Presidente della Repubblica Italiana?* una risposta come *Il neo presidente della Repubblica, Giorgio Napolitano, è giunto al Quirinale per incontrare il presidente della Repubblica uscente Carlo Azeglio Ciampi* sarebbe da considerarsi corretta e sufficientemente esatta, in quanto contenente un certo numero di informazioni non richieste

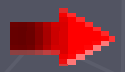
Cross Language QA (12/19)

- CLEF 2005 QA: risultati
  - Partecipazione
    - Il numero di gruppi partecipanti è stato abbastanza elevato (44 task attivati), a dimostrazione dell'interesse esistente per l'argomento
  - Risultati
    - 6 sistemi hanno raggiunto discrete prestazioni, superando la soglia del 40% di accuratezza delle risposte fornite alle 200 domande proposte dalla commissione valutatrice

Cross Language QA (13/19)

- Web CL Question Answering: la nuova frontiera
  - Problema
    - Nelle attività finora previste nell'ambito del CLEF, si è sempre prediletta come sorgente di conoscenza per i sistemi QA una collezione chiusa di documenti (un approccio ereditato dal TREC*), per motivi di controllo delle fonti e di semplicità di valutazione dei sistemi
    - La fonte di conoscenza più naturale (ed anche la più impegnativa) per un sistema CL QA è data dal Web…
  - Web come sorgente di conoscenza
    - Una mole di documenti enorme, disponibili per la ricerca di risposte a (quasi) ogni genere di domanda
    - Necessità di sistemi di traduzione efficienti

* Text Retrieval Evaluation Conference

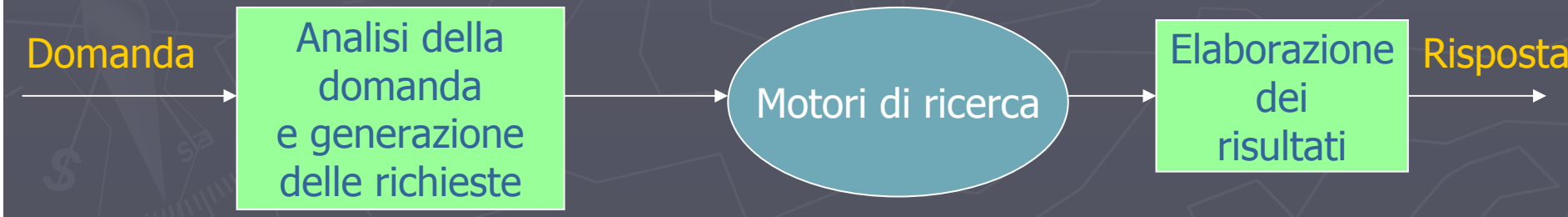


Cross Language QA (14/19)

- Web CLQA: il Web come sorgente di conoscenza
  - Vantaggi
    - Una quantità di informazioni estremamente superiore a qualsiasi corpus
    - Ridondanza delle informazioni, che sono tipicamente presenti più volte ed in forme differenti
    - Possibilità di applicare metodi statistici (se una informazione compare più volte in fonti differenti è meno probabile sia falsa o inesatta)
  - Svantaggi
    - Nessun controllo di qualità sulle fonti
    - Possibilità di errori grammaticali e di battitura, che complicano l'analisi che il sistema deve operare
    - Impossibile analizzare in anticipo i documenti (a causa della mutabilità del contenuto, ma soprattutto per via delle dimensioni improponibili)

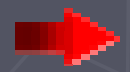
Cross Language QA (15/19)

- Web CLQA: metodologie proposte
  - Accesso mediante motori di ricerca
    - Si sfruttano indicizzazioni preesistenti dei documenti presenti sul Web (Google e Yahoo dichiarano di indicizzare oltre 18 bilioni di documenti ciascuno…)
    - Il controllo sulla ricerca può risultare limitato
    - Proposto come base di partenza per reperire una base documentale su cui effettuare ulteriori ricerche, con uno schema del tipo seguente:



Cross Language QA (16/19)

- Web CLQA: metodologie proposte
  - Sfruttamento della ridondanza delle informazioni
    - La presenza delle medesime informazioni espresse in varie forme rende più probabile trovare una risposta che sia una riformulazione della domanda
    - Non è necessario applicare tecniche troppo sofisticate di elaborazione del linguaggio naturale: per gran parte delle domande fattuali è infatti sufficiente applicare tecniche di parsing superficiale ai documenti restituiti dai motori di ricerca
    - Esempio: a partire dalla domanda *Chi ha ucciso Abramo Lincoln?* possiamo aspettarci di reperire molti documenti, di cui uno probabilmente conterrà la riformulazione *J. W. Booth uccise Abramo Lincoln*

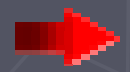


Cross Language QA (17/19)

- Web CLQA: metodologie proposte
  - Sfruttamento della ridondanza delle informazioni
    - L'ipotesi perde però di validità nel caso di domande complesse (*Quali furono le motivazioni politiche dell'esilio di Dante?*)
    - In tali casi si rende necessaria una forma di analisi più dettagliata, che permetta di classificare la domanda come appartenente ad una o più categorie (temporale, modale…) ed estrarre parole chiave, al fine di applicare tecniche più sofisticate
  - Trasformazione domanda ed estrazione parole chiave
    - Si tratta di tecniche applicate nel caso di domande complesse, per produrre una serie di richieste per i motori di ricerca

Cross Language QA (18/19)

- Web CLQA: metodologie proposte
  - Trasformazione domanda ed estrazione parole chiave
    - Trasformazione: la domanda è analizzata per produrre nuove richiesta mediante eliminazione di termini o modifiche nell'ordine delle frasi (*Chi è stato il primo uomo ad andare nello spazio?* → "Il primo uomo ad andare nello spazio è stato", "è stato il primo uomo ad andare nello spazio")
    - Estrazione delle parole chiave: dalla domanda si produce un insieme di parole chiave, così da comporre le richieste per i motori di ricerca mediante delle combinazioni di tali parole



Cross Language QA (19/19)

- Web CLQA: metodologie proposte
  - Traduzione della domanda VS traduzione delle richieste
    - La traduzione diretta della domanda in più lingue è certamente la strada più semplice, ma nel caso di domande complesse può portare ad errori di traduzione tali da invalidare la ricerca
    - La traduzione delle richieste generate, benché soggetta anch'essa ad errori di traduzione, lascia maggiori probabilità di ottenere risultati corretti, poiché una o più richieste possono essere in una forma di facile traduzione

Interactive CLIR (1/5)

- Interactive Cross Language Information Retrieval
  - Motivazioni
    - Una definizione critica del CLIR: “Il problema di reperire documenti che non puoi leggere, o peggio, che neanche sei in grado di riconoscere”
    - Nonostante la ricerca in questo campo vada avanti già da diversi anni e nonostante i buoni risultati ottenuti, non esistono nella pratica vere e proprie soluzioni commerciali
    - Diversi motori di ricerca offrono servizi di traduzione automatica delle pagine, ma non esiste nessun servizio, se non a livello sperimentale, di ricerca cross language
    - Gli attuali sistemi di CLIR interagiscono troppo poco con l’utente, e sono perciò poco appetibili a livello di mercato

Interactive CLIR (2/5)

- Interactive CLIR: tipologie di interazione
  - Interazione strettamente monolingua
    - Da utilizzare in tutti i casi in cui l'utente non ha alcuna conoscenza delle altre lingue in cui sono espressi i documenti (ad esempio un utente europeo che ricerchi informazioni in una collezioni di documenti giapponesi)
    - E' necessaria una traduzione completa sia in fase di selezione dei documenti, sia in fase di raffinamento della richiesta (traduzione completa per i documenti, tecnica di ritraduzione della query)
  - Interazione con utenza avente conoscenza passiva
    - Un utente con una conoscenza passiva di altre lingue, pur non avendo le capacità di generare richieste in tali lingue, è in grado di percepire errori di traduzione grossolani e di indicarli al sistema

Interactive CLIR (3/5)

- Un esempio d'applicazione: Interactive CL QA
  - Un modello meno restrittivo
    - Deviazione dal modello classico del QA: il sistema produce una o più risposte alla domanda, fornendo però anche i documenti di provenienza (eventualmente tradotti in maniera automatica). La possibilità di osservare il contesto da cui è scaturita la risposta consente all'utente di verificarne l'esattezza (perlomeno nei casi di conoscenza passiva della lingua o di traduzioni automatiche sufficientemente accurate)
  - Inserimento di informazioni sulla risposta attesa
    - Anche se l'utente non conosce la risposta alla domanda è ragionevole che sappia che risposta aspettarsi, specie nel caso di domande fattuali (la risposta sarà una data, un luogo, un nome proprio…)
    - Il sistema sfrutta tale conoscenza per ridurre il volume della ricerca, considerando solo risposte del tipo indicato

Interactive CLIR (4/5)

- ICL QA: un approccio alternativo
  - Utilizzo di sunti automatici
    - Piuttosto che tradurre interi documenti e costringere l'utente a lunghe analisi, gli si propongono dei brevi riassunti dei documenti di provenienza delle risposte
    - I riassunti possono essere di tipo Keyword in Context o single passage
  - Tipi di riassunto
    - Keyword in Context: si tratta di un riassunto composto di tre sole frasi, che devono tassativamente contenere almeno una delle parole chiave estratte dalla domanda di partenza o da una successiva espansione della richiesta. Si tratta di un riassunto di tipo indicativo, in quanto fornisce soltanto un'idea del contesto del documento

Interactive CLIR (5/5)

- ICL QA: un approccio alternativo
  - Tipi di riassunto
    - Single passage: la distribuzione delle parole chiave all'interno del documento viene utilizzata allo scopo di individuare passaggi significativi, che vengono poi estesi ai limiti di paragrafo più prossimi (nel caso questi non siano individuali si utilizza una finestra di dimensione prefissata). Qualora vengano individuati più passaggi con punti di sovrapposizione, questi vengono uniti in un unico passaggio. Per ogni documento viene selezionato il passaggio avente punteggio più elevato (determinato in base alla densità delle parole chiave). Si tratta di un riassunto informativo, in quanto è di norma sufficientemente ampio da permettere all'utente di stabilire buona probabilità la correttezza o meno della risposta fornita

CLIR: risorse utili

- Cross Language Evaluation Forum (CLEF): forum di valutazione di gran parte delle ricerche a livello internazionale sulle applicazioni del CLIR [http://www.clef-campaign.org]
- Text Retrieval Conference (TREC): manifestazione annuale, sponsorizzata dal NIST, su ricerche e applicazioni nel campo dell'IR, ed in particolare del text retrieval [http://trec.nist.gov]
- European Chapter of the Association for Computational Linguistics (EACL): conferenza triennale sui progressi nel campo Della linguistica computazionale [http://www.eacl.org]
- ACM SIGIR [http://www.sigir.org]

Bibliografia (1/3)

- D. W. Oard, A. R. Diekema – *Cross Language Information Retrieval* [Annual Review of Information Science and Technology, vol.33, 1998]
- F. C. Gey, N. Kando, C. Peters – *Cross Language Information Retrieval: the Way Ahead* [Information Processing and Management, Volume 41]
- M. L. Littman, S. T. Dumais, T. K. Landauer – *Automatic CLIR using Latent Semantic Indexing* [Grefenstette, G., editor, Cross Language Information Retrieval. Kluwer, 1998]
- T. A. Letsche, M. W. Berry - *Large-Scale Information Retrieval with Latent Semantic Indexing* [Information Sciences, Applications, 1997]
- Yih-Chen Chang, Wen-Cheng Lin, Hsin-Hsi Chen – *Combining Text and Image Queries at ImageCLEF2005*
- GNU Image Finding Tool (GIFT) [*www.gnu.org/software/gift*]
- F. Long, H. Zang, D. D. Feng – *Fundamentals of Content Based Image Retrieval*
- SMART Information Retrieval System [*ftp://ftp.cs.cornell.edu/pub/smart/*]

Bibliografia (2/3)

- M. E. Ruiz, S. B. Southwick – *UB at CLEF 2005: Medical Image Retrieval Task*
- A. R. Aronson - *Effective Mapping of Biomedial Text to the UMLS Metathesaurus: The Metamap Program* [Proc AMIA 2001]
- Unified Medical Language System [http://www.nlm.nih.gov/research/umls]
- G. Salton – *The SMART Retrieval System: Experiments in Automatic Document Processing* [Prentice Hall, 1971]
- A. Vallin, D. Giampiccolo, B. Magnini – *Overview of the Multilingual Question Answering Track at CLEF 2005*
- J. Gonzalo, D. W. Oard – *iCLEF 2004 Track Overview: Pilot Experiments in Interactive Cross Language Question Answering* [Springer, LNCS vol.3491, 2005]
- Wikipedia – Definizione QA [http://en.wikipedia.org/wiki/Question_answering]
- B. Magnini, A. Vallin et al. – *Overview of the CLEF 2004 Multilingual Question Answering Track* [Springer, LNCS vol.3491, 2005]
- J. Lin, B. Katz – *Question Answering Techniques for the World Wide Web* [EACL 11° Conference Proceedings]

Bibliografia (3/3)

- D. W. Oard – *Interactive Cross Language Information Retrieval*
- J. Gonzalo – *Scenarios for Interactive Cross Language Retrieval Systems*
- J. Gonzalo, F. Lopez-Ostenero, F. Verdejo, V, Peinado – *Interactive Cross Language Question Answering: Searching Passages Versus Searching Documents*
- D. He, J. Wang, J. Luo, D. W. Oard – *iCLEF 2004 at Maryland: Summarization Design for Interactive Cross Language Question Answering*