Il CLEF [Cross-Language Evaluation Forum], l'iniziativa europea per la valutazione dei sistemi di *information retrieval* nei contesti multilingui: una messa a punto

La rivista della DGI [Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis], "Information Wissenschaft & Praxis – Competence in Content", sul numero di marzo di quest'anno (2002, n. 2, p. 82-89) ha pubblicato uno studio particolareggiato delle iniziative in corso in questo campo.

Gli autori (Michael Kluck, Thomas Mandl e Christa Womser-Hacker) ne offrono la rassegna completa, collegando l'ambizioso progetto europeo con analoghe esperienze condotte in Nordamerica e in Giappone. E corredano l'articolo di una ricca ed agile bibliografia (p. 89), in larghissima parte in lingua inglese e dunque rivolta ad un pubblico internazionale, che informa puntualmente sullo stato dell'arte.

Con l'avvento di Internet, l'IR ha acquistato una straordinaria importanza: grandi quantità di conoscenze, immagazzinate e accessibili *online*, sono divenute in ampia misura liberamente fruibili da parte degli utenti tramite i motori di ricerca.

Parallelamente, è cresciuta ovunque nel mondo l'esigenza di valutare, secondo criteri standard, i sistemi, appunto, di IR.

Ai problemi generali che tradizionalmente lo accompagnano si aggiungono, nell'**IR multilingue**, quelli difficili e spinosi della traduzione e della presentazione integrata dei risultati, a partire da grandi quantità di documenti.

Il **Crosslingual IR** (CLIR), dal canto suo, tenta di fornire, sulla base della richiesta formulata in una lingua, documenti in altre lingue e va in cerca di documenti rilevanti in un *corpus* multilingue.

Iniziative internazionali finalizzate alla valutazione

Dato che inizialmente i ricercatori utilizzavano raccolte di testi assai diversificate, gli esiti della ricerca risultavano scarsamente confrontabili.

Negli ultimi anni, però, sono state messe a disposizione collezioni allestite secondo standard comuni ed è così migliorata la confrontabilità fra i sistemi.

Proprio lo sviluppo dei metodi di *retrieval* e di un'adeguata infrastruttura per la valutazione di questi stessi metodi di ricerca, impiegati per il superamento delle barriere linguistiche, è lo scopo del CLEF.

Nato tre anni fa, si basa su un'esperienza pilota statunitense, il **TREC** [Text Retrieval Conference].

In Giappone, contemporaneamente al CLEF, è sorto l'**NTCIR** [NII-NACSIS Test Collection for IR Systems], per la ricerca multilingue nell'àmbito delle lingue asiatiche.

Il TREC [Text Retrieval Conference]

Negli USA dal 1989 il NIST [National Institute of Standards and Technology] porta avanti un progetto per la valutazione dei sistemi di IR. Mette pertanto a disposizione un'enorme massa di dati, i *topic* e l'infrastruttura per la valutazione.

A motivo dell'ampio consenso riscosso dall'iniziativa, numerosissimi sono i gruppi di ricerca, attivi nel settore industriale e in quello scientifico, che con i loro sistemi partecipano annualmente al TREC. I risultati del TREC 2001 sono disponibili *online* (<http://www.clef-campaign.org>).

Sul piano operativo, il TREC è suddiviso in vari gruppi di lavoro (*tracks*), ciascuno impegnato ad indagare e, se possibile, a risolvere un singolo aspetto della complessa problematica.

Quest'anno s'è aggiunto il *Video retrieval track* e il *Web track* – che fornisce l'istantanea aggiornata di un segmento di Internet – ha preso il posto dell'*Ad-hoc-retrieval track*.

Se la sua versione ridotta consta di 1,7 milioni di siti (10 Gigabyte), quella ampia ne abbraccia ben 18,5 milioni (100 Gigabyte).

Nel 1994 il TREC ha avviato il *Cross-language track*, comprendente dapprima solo documenti in inglese e in spagnolo; successivamente si sono aggiunti il cinese e, dal 1997, le lingue europee e l'arabo. L'esperimento, senza molta fortuna negli Stati Uniti, è proseguito in Europa, al punto che il CLEF se ne può considerare il legittimo erede.

Il CLEF [Cross-Language Evaluation Forum]

Il CLEF mette a frutto, come s'è detto, l'esperienza maturata dal TREC nel *Cross-language track* riservato alle lingue europee.

Dotato di una struttura articolata – IEI-CNR (Pisa, Italia), cui è affidato il coordinamento, Eurospider (Zurigo, Svizzera), ELRA (Parigi, Francia), IZ (Bonn, Germania), UNED (Madrid, Spagna), NIST (Gaithersburg, USA) –, si vale del lavoro di gruppi provenienti da vari Paesi europei (vale a dire dalle corrispondenti aree linguistiche) e collabora attivamente con il NIST.

I *topic* per la formulazione delle domande dei test sono predisposti su tre livelli descritti in modo dettagliato. Accanto ad un **titolo** (*title*), costituito di poche parole, vi è una **breve descrizione** (*description*) ed una cosiddetta "**descrizione lunga**" (*narrative*).

I partecipanti optano per una sola delle tre formulazioni o per una loro combinazione (per esempio, *title* e *description* ovvero tutti e tre gli elementi insieme).

Il CLEF promuove la ricerca e lo sviluppo dell'IR *crosslingual* e *multilingual* attraverso la messa a punto di un'infrastruttura che sia a disposizione per:

- i test cui sottoporre i sistemi di IR;
- la valutazione dei sistemi di IR applicati alle lingue europee;
- la produzione di *testsuite* di dati che possano essere ancora utilizzati dagli sviluppatori di sistemi per il *benchmarking*.

Contestualmente, dovrebbe perciò sorgere un *forum* di discussione per scambiare esperienze ed idee e per favorire la comunicazione fra scienza ed economia nel campo del CLIR. E dovrebbe inoltre essere agevolato il trasferimento di tecnologie fra gli istituti di ricerca e gli utenti commerciali.

AMARYLLIS

L'organizzazione del progetto AMARYLLIS, relativo alla lingua francese, è affidata all'INIST-CNRS; si adegua, dal punto di vista metodologico, ai principî del TREC. Dopo due fasi dedicate esclusivamente al francese (1996-97 e 1998-99), attualmente partecipa al CLEF.

Metodica di valutazione

La sperimentazione si fonda, per le diverse lingue, su raccolte parziali di articoli di giornale e di comunicati di agenzie di stampa.

Sono a disposizione le seguenti.

Giornali e comunicati di agenzie di stampa:

- inglese – 113.005 documenti, 425 MB;
- tedesco – 225.371 documenti, 527 MB;
- francese – 87.191 documenti, 243 MB;
- italiano – 108.578 documenti, 278 MB;
- spagnolo – 215.738 documenti, 509 MB.

Dati scientifici e di àmbito specialistico:

- scienze (tutti i settori): AMARYLLIS (francese) – 150.000 documenti, 20 MB;
- scienze sociali: GIRT [German Indexing and Retrieval Database] (tedesco) – 76.128 documenti, 150 MB.

Dati supplementari per test bilingui (giornali e comunicati di agenzie di stampa):

- olandese – 190.604 documenti, 540 MB.

Le raccolte sono obbligatoriamente complete per il 1994 e abbracciano in parte anche il 1995.

I singoli documenti sono provvisti dei *tag* SGML.

È imminente l'integrazione di un *corpus* finlandese e di uno svedese.

Nel lungo periodo, inoltre, è molto atteso l'allargamento alle lingue dell'Europa orientale.

La creazione dei soggetti

I soggetti sono creati dai diversi gruppi linguistici del CLEF (tedesco, inglese, spagnolo, francese, italiano): debbono ovviamente corrispondere al contenuto dei documenti, sulla base dei giornali e dei comunicati delle agenzie di stampa per gli anni 1994-95.

I gruppi linguistici derivano mappe (le "estraggono", vale a dire le rintracciano retrospettivamente sulla scorta di annuari ed enciclopedie) relative a questo periodo. Successivamente, testano i concetti espressi dai soggetti sugli insiemi di dati nelle rispettive lingue.

Il sistema ZPRISE – con cui si effettuano i test preliminari – conduce una ricerca probabilistica e contiene un *feedback* di rilevanza, il quale permette di contrassegnare i documenti rilevanti e di fornire al sistema tale informazione aggiuntiva; valuta poi le probabilità sulla base dell'ultimo *feedback* e rintraccia automaticamente i concetti supplementari, che vengono quindi incorporati nella domanda.

La decisione definitiva circa la scelta dei soggetti e il numero delle mappe è il frutto di una discussione collettiva. Questa ha lo scopo di chiarire ai partecipanti l'effettivo valore delle mappe e delle traduzioni nelle rispettive lingue.

È indispensabile, perciò, sottoporre ad un controllo finale le traduzioni nelle 5 lingue ufficiali, per garantirne l'efficacia e l'esattezza.

L'intero processo si compie in modo cooperativo.

Per le mappe dei problemi scientifici e di àmbito specialistico sono già sviluppati, nel quadro del GIRT (<http://www.iei.pi.cnr.it/DELOS/CLEF/clefoo.html>) e dell'AMARYLLIS, 25 soggetti specifici per disciplina, in tedesco e in francese; esiste di essi una traduzione inglese (e nel GIRT anche una russa), per permettere i test del CLIR.

Sono altresì pronte, presentate ufficialmente dai corrispondenti gruppi linguistici, le mappe dei soggetti in inglese, tedesco, francese, italiano, spagnolo, olandese e russo (GIRT). E, per l'IR monolingue e bilingue, traduzioni non ufficiali di tutte le mappe sono già state predisposte dai partecipanti per il finlandese, il greco, lo svedese, il russo, il cinese, il tailandese e il giapponese.

Check* dei *topic

Per escludere ogni specie di errori, il *topic-set* definitivo è sottoposto ad una verifica cruciale, affidata ad un gruppo indipendente di traduttori specialisti, plurilingui e dotati di competenze interculturali.

Percentualmente, il numero più elevato di errori è costituito dalle deviazioni dal testo di partenza e dagli errori stilistici e grammaticali; il più basso dagli errori di ortografia e di interpunzione.

Le lingue

Per "lingua principale" all'interno del CLEF s'intende una lingua in cui siano presenti e disponibili una o più raccolte di documenti e tutti i soggetti e le relative mappe.

Al momento, sono le seguenti: tedesco (DE), inglese (EN), spagnolo (ES), francese (FR), italiano (IT).

Sono appunto le lingue principali a ricoprire il ruolo decisivo nell'IR multilingue: fra queste, infatti, gli sviluppatori dei sistemi scelgono quale debba costituire il punto di partenza per le loro ricerche.

Nel 2001 si sono aggiunti: finlandese (FI), olandese (NL), russo (RU), svedese (SV), tailandese (TH), giapponese (JP), cinese (ZH).

Questioni aperte

L'obiettivo primario del CLEF è, come s'è detto, lo sviluppo dell'**IR multilingue** (*multilingual task*): vale a dire la ricerca nei documenti di tutte le lingue principali – in una di esse è formulata la *query* – e l'*output* di un elenco integrato di tutti i risultati, provenienti da tutte le raccolte dei documenti (cioè da tutte le lingue principali).

Si pensa di poter utilizzare, a breve, altre lingue (finlandese, russo, svedese) come lingue di partenza, mentre le lingue di destinazione rimarrebbero quelle principali.

L'**IR bilingue** (*bilingual task*) cerca, in una qualsiasi lingua di partenza (che non sia uguale a quella di destinazione), documenti in olandese e in inglese.

Al contrario, **quello monolingue** (*monolingual task*) punta a rintracciare documenti all'interno delle rispettive raccolte, in tedesco, inglese, francese, olandese, italiano e spagnolo. Il *monolingual task* offre l'opportunità di allargare il progetto a nuove lingue, che in questo modo saranno progressivamente integrate nella sperimentazione dell'IR multilingue.

Per quel che riguarda **il campo scientifico in senso lato e specifici settori** (*scientific and domain-specific task*), si cercano documenti scientifici (e, più precisamente, relativi alle scienze sociali) in particolari collezioni di documenti, vale a dire il GIRT e l'AMARYLLIS.

I documenti delle loro banche dati contengono altresì parole d'ordine assegnate di volta in volta, *a posteriori* e in modo ragionato, a un tesauro scientifico (e, più esattamente, di scienze sociali), tesauro disponibile pure in una traduzione inglese e, limitatamente al GIRT, in una russa.

Specifiche mappe dei soggetti sono pronte in inglese, tedesco, francese e russo (GIRT).

L'AMARYLLIS e il GIRT, dunque, forniscono una piattaforma ideale per testare la trasferibilità dei sistemi di IR su testi di particolari àmbiti scientifici.

Ancora del tutto sperimentale è invece la ricerca nel campo del **CLIR interattivo** (*interactive task*): questo *track* mira a studiarne la valutazione e a sviluppare misure e criteri confrontabili, con cui possano essere comparati studi successivi.

Assai promettente appare infine la possibilità di formulare e di variare la domanda calibrandola via via, valutando rapidamente i documenti trovati. In tal caso, le domande sono rielaborate dalle persone che partecipano al test e non poste automaticamente dal sistema o dagli esperti.

Revisione delle mappe dei soggetti ad opera dei partecipanti al progetto

I sistemi di *retrieval* cercano le mappe dei soggetti in una lingua e restituiscono documenti in tutte le lingue di destinazione.

Per le ricerche in raccolte di documenti (nell'IR multilingue: tedesco, inglese, spagnolo, francese, italiano) impiegano strategie specifiche per sistema, con l'obiettivo di superare le difficoltà inerenti alla traduzione e alla trasformazione delle domande in lingue differenti.

Alla fine dei processi di estrazione delle risposte, dovrebbe risultare, in relazione alle mappe dei soggetti, una serie comune e ordinata dei primi 60 documenti rilevanti.

Accanto alla problematica della traduzione, la sfida fondamentale è rappresentata dal processo d'integrazione dei risultati, sulla base di insiemi di documenti diversi.

Il metodo di valutazione si fonda sul **metodo Pooling** del TREC: al termine dei processi, si raccoglie un numero elevato di documenti in serie di risultati ripartiti secondo le lingue per mappe di soggetti.

La valutazione della rilevanza

Questi elenchi di risultati sono poi valutati dai membri della giuria del rispettivo gruppo linguistico, i quali si servono di un apposito software di valutazione, ASSESS, sviluppato dal NIST. Le regole generali di valutazione, cui si attengono, sono confrontabili con quelle del TREC.

Le discussioni dei gruppi linguistici relative ai soggetti costituiscono le linee guida per i loro giudizi che si appoggiano principalmente sulle descrizioni lunghe (*narrative*).

Pur manifestando spesso il desiderio di una scala graduata di rilevanza, i membri della giuria si attengono al pur difficile giudizio binario. Così si regola infatti il CLEF, al pari del TREC, per privilegiare una migliore utilizzabilità.

In séguito ad un definitivo riordinamento dei risultati complessivi per sistema e per soggetto, sono prodotte curve di richiamo-precisione, in corrispondenza di ciascun sistema e del confronto fra i sistemi.

I *trend* nel *retrieval* multilingue

Com'è noto, la problematica cruciale del *retrieval* multilingue è il trattamento dell'eterogeneità.

I procedimenti fondamentali si possono suddividere in tre gruppi:

- traduzione delle *query*;
- traduzione di tutti i documenti;
- metodi associativi senza traduzione esplicita.

I sistemi, inoltre, si distinguono in base al modo di elaborazione linguistica:

- riduzione alle forme originarie (*stemming*);
- scioglimento dei composti (*decomposition*);
- parole o n-grammi come fondamento.

Il nucleo delle ricerche multilingui è costituito dal trasferimento delle *query* o delle rappresentazioni dei documenti dalla lingua di partenza a quella di destinazione.

La traduzione di tutti i documenti, che in passato rappresentava il maggior ostacolo a motivo degli alti costi, oggi si realizza facilmente grazie ai moderni computer. Numerosi sono i mezzi impiegati a tale scopo, dai più diffusi in commercio a quelli liberamente accessibili in Internet.

Per garantire al sistema più punti di appoggio per la rilevanza, nella lingua di destinazione le domande sono spesso ampliate; e si associano pure termini aggiuntivi, semanticamente affini.

Questi sono definiti sulla base di dizionari e di tesauri ovvero delle occorrenze statistiche in un *corpus*.

Un'altra possibilità è offerta dai procedimenti sfumati, associativi, che rinunciano a rapporti sicuri, così come sono conosciuti dai dizionari.

Tali sistemi apprendono, mediante metodi automatici, le relazioni tra le parole nelle diverse lingue. Pertanto hanno bisogno di un doppio *corpus*, cioè di documenti identici in entrambe le lingue per la quantità di *training*.

Divengono quindi decisive le associazioni di concetti semanticamente simili. Nasce così un tesauro delle somiglianze tra due lingue, che si fonda su giacimenti comuni.

I sistemi presentati al CLEF mostrano i seguenti *trend*.

1. I sistemi associativi e basati sul *corpus* crescono d'importanza, sia per la traduzione che per la disambiguazione.
2. Non sono state studiate a sufficienza le conseguenze delle operazioni di base, quali la riduzione alle forme originarie e la scomposizione dei composti;

in tedesco e in olandese, come pure in finlandese e in svedese, la scelta dell'algoritmo ha condizionato fortemente i risultati.
Quanto allo scioglimento dei composti, i risultati sono assai divergenti: grazie a questo, alcuni gruppi hanno ottenuto considerevoli miglioramenti, altri invece non ne hanno ricavato alcun profitto.

3. Un certo successo riscuotono le tecniche fondate sugli n-grammi, che non si valgono di alcun modello linguistico – lo sviluppatore delle quali non possiede, anzi, alcuna conoscenza linguistica. Ciò riesce tanto più significativo, se paragonato ai sistemi maggiormente sofisticati, con elementi, cioè, più perfezionati dal punto di vista linguistico.

Nel Workshop 2001 del CLEF a Darmstadt è apparso chiaro che i partecipanti sostituiscono largamente i componenti l'uno con l'altro, cosicché i sistemi sono parzialmente costituiti da moduli eterogenei.

Al contrario, nel campo della semantica s'è fatto ben poco. Nessun gruppo, infatti, ha provato ad analizzare l'attribuzione di valori negativi ai soggetti.

Per la prima volta, s'è anche proposto d'introdurre nel CLEF la problematica relativa alle lingue parlate; ciò dovrebbe suscitare l'interesse dei ricercatori nel settore dell'identificazione linguistica.

Conclusioni

Il progetto riesce assai utile per il collaudo dei sistemi su dati reali, ben oltre l'*information retrieval*:

- l'àmbito multilingue può rivelarsi decisivo per i ricercatori impegnati nei settori della linguistica computazionale e della traduzione automatica;
- per il successo dei sistemi di *retrieval*, la morfologia resta un elemento cardine, ma pure competenze relative alla sintassi e alla semantica offrono prospettive interessanti;
- la sperimentazione condotta all'interno del *track* interattivo è un'opportunità da non sottovalutare per quanti si occupano dell'interazione uomo-macchina.

Il Workshop 2002, infine, si svolgerà a Roma (19-20 settembre 2002) e terrà immediatamente dietro alla sesta edizione della European Conference on Digital Libraries (ECDL 2002). Per il programma, assai ricco e articolato, si veda: <http://clef.iei.pi.cnr.it:2002/2002work.html>.