Introduzione

I rapidissimi sviluppi delle tecnologie legate ai personal computer hanno portato negli ultimi anni ad una esplosione della quantità di informazione pubblicata su Internet e sul world wide web in particolare. Purtroppo però, proprio a causa della dimensione del fenomeno, non è umanamente possibile fruirne completamente e per il futuro la situazione non può che peggiorare.

Tale problema è di centrale importanza per chi unque debba prendere delle decisioni, in particolare sotto i prifili del “cosa fare” e del “momento giusto”. Si pensi per esempio a chi deve acquistare una vettura o un immobile, o al manager che deve decidere quando immettere sul mercato un elettrodomestico rivoluzionario o a chi, più banalmente, vuole acquistare di un telefonino o un lettore mp3. Ovviamente egli deve seguire attentamente quello che accade sul mercato facendo particolare attenzione alle notizie che lo riguardano più da vicino. Considerando l'enorme quantità di fonti di informazione è praticamente impossibile riuscire a scoprire ed acquisire tutti gli annunci, le notifiche, le segnalazioni i commenti e quant'altro interessi o, in altre parole, tutta la conoscenza pertinente.

Per questi ed altri scopi sarebbe di estrema utilità disporre di un sistema che in modo automatico individui quali sono le informazioni rilevanti e fino a quel momento sconosciute, rendendole fruibili in modo semplice.

Per questi ed altri scopi è nata la Topic Detection and Tracking (TDT) per affrontare e risolvere il problema di identificare delle storie, all'interno di flussi di informazione, pertinenti ad un certo evento.

Si parla di flusso di informazione in quanto l’ambizioso obiettivo finale è analizzare ogni possibile fonte di informazione, sia essa carta scritta, il web in tutte le sue sfaccettature, notiziari televisivi o radiofonici, ecc, il tutto in ogni possibile lingua. Viene assunto che ognuna di queste fonti emetta informazione sotto forma di una sequenza di storie separate tra loro, ognuna inerente a uno o più eventi.

Una prima grossolana definizione di evento potrebbe essere “qualcosa che accade in un ben preciso momento”. Volendo fare un esempio, “i mondiali di calcio 2006 in Germania” sono un evento, mentre “i mondiali di calcio” sono una classe di eventi, a sua volta sottoclasse di “sport”. Inoltre, esso può essere previsto, come un referendum o delle elezioni politiche, o inatteso, come un terremoto o uno scudetto dell'Inter.

In generale la Topic Detection and Tracking si occupa di identificare qual’è l’evento discusso in queste storie o quali storie parlino di un determinato evento.

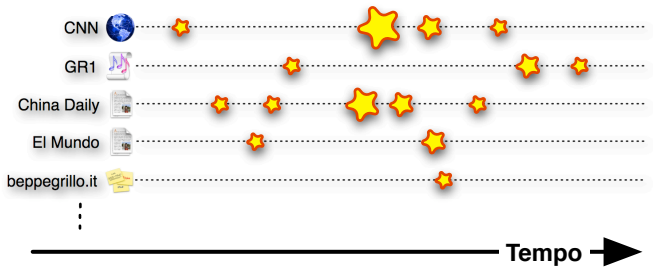

Figura 1: Storie che fanno riferimento ad un evento in molteplici fonti

La figura 1 illustra una tipica situazione: diversi media che nel tempo trattano un certo evento una o più volte: il notiziario televisivo della CNN, il radiogiornale GR1, il sito della testata giornalistica cinese China Daily, il quotidiano El Mundo, il blog di Beppe Grillo.

La TDT si prefigge l’obiettivo di scoprire una simile struttura in maniera automatica.

Eventi, attività, argomenti e storie

Un *evento* (in inglese “TDT event”) è un particolare avvenimento che succede in un preciso momento ed in un preciso luogo, assieme alle sue cause e le inevitabili conseguenze.

Le *attività* (in inglese “TDT activity”) sono insiemi di eventi con lo stesso scopo, che accadono in ben determinati momenti e luoghi.

Un *argomento* (traduzione molto sfortunata dell’inglese “TDT topic”) è definito come un evento capostipite (altra traduzione sfortunata dall’inglese “seminal event”) o una attività, unitamente a tutti gli eventi ed attività strettamente collegati.

Ci si riferisce spesso ad una *storia* intendendo un articolo di giornale, una trasmissione radiofonica, televisiva o quant’altro rechi informazioni su un determinato evento.

Queste definizioni sono piuttosto ambigue: un TDT topic infatti è un concetto diverso da quello dell’usuale topic. Il secondo infatti è un generico argomento come potrebbe essere la politica o lo sport, nell’uso comune

ha quindi un significato molto più ampio rispetto a quello inteso all'interno della Topic Detection.

Per capire meglio si consideri l'esempio:

Nel febbraio 1998 un jet della marina militare americana volando a bassa quota ha tranciato i cavi di supporto della funivia che da Cavalese porta sul monte Cermis. Il veicolo è entrato in collisione con le funi quando la cabina passeggeri si trovava a 300 metri dalla stazione di arrivo, facendola precipitare da circa 110 metri d'altezza provocando la morte di tutti e 20 i passeggeri.

Lo schianto della cabina e le conseguenti morti sono eventi i quali sono dirette conseguenze della pessima scelta di traiettoria del jet e perciò sono considerati come parte dello stesso evento capostipite.

Per passare dal concetto di evento a quello di topic bisogna allentare la concentrazione sul fatto "strage del Cermis" e considerare per esempio gli sforzi che fecero le unità di soccorso, i funerali delle vittime, le dichiarazioni che fece la Marina a proposito delle strategie di addestramento in aree civili, le indagini di polizia che seguirono, l'incredibile sentenza della Corte Marziale che scagionò completamente il pilota.

Questi sono eventi direttamente collegati al capostipite e quindi considerati facenti parte dello stesso topic; si dicono essere "in topic".

L'aereo che ha causato la strage faceva parte delle forze armate che sorvolavano la Bosnia durante la guerra. Questo fatto, tuttavia, non è sufficiente a creare un collegamento tra gli avvenimenti di Cavalese e la guerra in Bosnia, che in questo caso si dice essere "off topic", cioè che non c'entra nulla.

C'è da aggiungere che queste definizioni hanno subito profondi cambiamenti nel corso degli anni a causa dell'approccio che la comunità scientifica ha adottato nell'affrontare il problema.

Storia del TDT Evaluation Project

L'idea ha visto la luce nel 1996, quando la DARPA (Defence Advanced Research Projects Agency, ente americano che da decenni finanzia grosse ricerche, tra cui la più nota è certamente quella che ha prodotto Internet) ha sentito il bisogno di possedere una tecnologia in grado di riconoscere la struttura delle notizie in flussi di informazione senza l'intervento umano.

Nel 1997 un primo studio getta le basi stimando fattibilità e convenienza ottenibili dalla nascita di un programma di ricerca e valutazione: un sistema per il quale vengono decisi dei task (in pratica delle specie di sotto problemi, come si vedrà più avanti) da perseguire e dei metodi di valutazione dei risultati. Il sistema prevede la creazione di un insieme di dati (un corpus) comune e disponibile a tutti, sul quale provare gli algoritmi secondo delle specifiche modalità di test.

Il programma nasce prendendo in esame solamente il formato testuale. Viene costruito un piccolo corpus denominato TDT-pilot di 15863 documenti contenenti notizie fornite dalla Reuters North American e dalla CNN, separate manualmente. Lo scopo è riconoscere gli eventi legati a 25 topic scelti appunto come i target della valutazione. La Carnegie Mellon University e la University of Massachussets sono le prime a partecipare con dei propri progetti.

Nel 1998 viene bandito dal NIST (National Institute of Standards and Technology) il "Topic Detection and Tracking Project 1998" , un progetto di valutazione delle tecnologie legate alla TDT.

Per l'occasione il Linguistic Data Consortium (LDC) produce, in collaborazione con Dragon (azienda leader nel settore del riconoscimento vocale), un corpus più eterogeneo e decisamente più sostanzioso del precedente, che prende il nome di TDT2. I 60000 documenti presenti includono articoli del New York Times e della Associated Press Worldstream, trascrizioni automatiche di circa 800 ore di trasmissioni tra quelle radiofoniche di Voice Of America e Public Radio International e televisive della CNN ed ABC. Il corpus viene dettagliatamente segmentato ed annotato dal LDC che seleziona 100 topic, più o meno casuali, come target della valutazione.

Ai partecipanti della precedente edizione si aggiungono la University of Pennsylvania e colossi come la General Electric ed IBM.

Nel 1999 viene riproposta ai ricercatori di tutto il mondo la valutazione delle tecnologie di topic detection, vista però più come una estensione di quella svoltasi nel 1998 che come una nuova manifestazione.

Il corpus della passata edizione viene ampliato includendo per la prima volta storie in una lingua diversa dall'inglese: il cinese. Inoltre ne viene creato uno nuovo, chiamato TDT3, attingendo alle fonti dell'anno precedente ed aggiungendoci materiale scritto in cinese fornito da Xinhua News e dal sito web "Zaobao", trascrizioni di programmi della radio Voice Of

America Mandarin, e trasmissioni televisive di NBC e MSNBC.

Il nuovo corpus contiene circa 63000 storie e definisce 60 topic target, questa volta aventi il requisito di essere trattati in almeno 4 storie per ognuna delle lingue presenti.

I corpus TDT2 e TDT3 assieme contengono circa 30000 storie in mandarino, cosa che attira ulteriori partecipanti quali la National Taiwan University e la MITRE, un ente americano per la ricerca su compiti particolarmente complessi.

Questa edizione modifica sostanzialmente alcune definizioni cardine all'interno dei criteri di valutazione degli algoritmi proposti ed aggiunge dei sotto compiti più precisi e raffinati (trattati nel dettaglio fra breve) per stimolare i ricercatori coinvolti.

Nel 2000 le redini dell'appuntamento fisso con la campagna di valutazione vengono riprese dalla DARPA, che la affianca al proprio programma TIDES (Translingual Information Detection and Summarization).

Questa edizione non estende i corpus esistenti ma sceglie solamente altri 60 eventi target. Vengono modificati ulteriormente sia i criteri di valutazione che i sotto compiti.

Il numero dei partecipanti non smette di crescere. Anche la Chinese University of Honk Kong e la TNO, organizzazione non-profit olandese per la ricerca, sottopongono le proprie idee alla valutazione.

L'edizione del 2001 non brilla di luce propria: non ci sono cambiamenti al corpus rispetto all'edizione precedente, eccezion fatta per la selezione di altri 60 eventi target sulla quale effettuare la valutazione.

Per quanto riguarda i partecipanti si segnalano le proposte francesi del Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur e l'uscita di qualche team, come Dragon, presente fin dalle prime edizioni.

In effetti questa edizione sembra concentrarsi più sul futuro che sul presente, proponendo per la prima volta l'inclusione nel corpus di storie in lingua araba. Gli argomenti della valutazione sono infatti più o meno gli stessi da qualche anno, e si sente il bisogno di nuovi stimoli alla ricerca.

Nel 2002 viene creato un nuovo corpus, il TDT4, prelevando storie da venti fonti in tre lingue: inglese, cinese ed arabo. China National Radio, China Central Television, Al-Hayat, Al-Nahar, Nile Tv solo per citarne alcune oltre alle già citate in lingua inglese.

Tra le 90735 storie vengono individuati 80 eventi trattati in tutte e tre le lingue su cui verranno effettuate le valutazioni.

Il sesto appuntamento annuale, quello del 2003, utilizza ancora il TDT4, focalizzando l'attenzione dei ricercatori sulla lingua araba. Vengono infatti aggiunti 40 nuovi topic target, scelti in modo da avere 10 eventi capostipite in lingua cinese, 10 in inglese e ben 20 in arabo.

Questo fatto mina le fondamenta dei risultati ottenuti negli anni precedenti provocando un deterioramento delle prestazioni e della precisione degli algoritmi proprio a causa dell'introduzione di una forte componente araba nel corpus e nei criteri di valutazione.

Il nuovo corpus contiene le registrazioni e le trascrizioni di circa 600 ore di trasmissioni televisive e radiofoniche nelle tre lingue. Sempre al fine di stimolare la ricerca per la prima volta le annotazioni in topic oppure off topic diventano strettamente binarie, in modo differente dai corpus precedenti in cui c'era anche l'opzione "brief" per indicare storie che trattavano anche solo parzialmente il topic in questione, utilizzata spesso in caso di dubbi.

Nel 2004 si svolge la settima ed ultima campagna di valutazione delle tecnologie legate alla topic detection, con delle importanti novità.

Viene innanzitutto creato un nuovo corpus, il TDT5, in cui, andando contro le tendenze seguite nei precedenti, sono presenti solo notizie in formato scritto, nessun programma radiofonico o televisivo. Il loro numero supera l'imponente cifra di 400.000 storie, obbligando di fatto i ricercatori a misurarsi con la complessità degli algoritmi proposti.

I topic scelti come target sono circa 250, divisi in 4 parti uguali, una per ogni lingua trattata più una parte di topic multilingua. Le annotazioni seguono delle direttive diverse, permettendo ora l'esistenza di topic incompleti, cioè con delle storie in tema non annotate come tali, e di topic trattati in una o in moltissime storie. E' possibile inoltre che certi argomenti si sovrappongano come informazioni contenute.

Quest'ultima osservazione è significativa in quanto una delle novità più importanti è che la ricerca viene focalizzata sulla rappresentarzione di topic in gerarchie di cluster piuttosto che in cluster distinti, permettendo delle sovrapposizioni, come si vedra' meglio tra breve.

Grazie a queste novità il numero di partecipanti è tornato ad essere cospicuo. Hanno partecipato infatti diverse università degli Stati Uniti e della Cina, più vari

enti ed aziende di tutto il mondo come l'olandese TNO e il cinese Institute of Computing Technology.

Come detto quella del 2004 è stata l'ultima campagna di valutazione. Il NIST ha infatti deciso di congelare il progetto visto che le tecnologie sviluppate durante gli ultimi anni non hanno introdotto novità significative rispetto alle loro cugine più anziane.

I task

Come anticipato, il paradigma di ricerca sulla Topic Detection è stato storicamente quello dello sviluppo e valutazione. Si fissano dei sotto compiti più semplici del problema principale, chiamati task, si definiscono le modalità di test e di valutazione dei risultati e si dà il via alla ricerca.

Dal 2000 in poi alcuni task sono classificati come primari. Questo significa che ogni squadra che ha intenzione di sottoporsi alla valutazione deve necessariamente trattare uno o più task primari ed uno di questi dev'essere quello su cui ci si concentra maggiormente. Questi task rappresentano problemi le cui soluzioni sono applicabili a diverse applicazioni di TDT e quindi le più utili.

★ Story segmentation task

Viene definito come il processo di suddivisione di un flusso di dati proveniente da una certa fonte in parti costituite da storie. Visto che il formato testuale è già diviso per natura questo task si applica unicamente ad un flusso in formato audio, cioè fonti come programmi radiofonici e/o televisivi.

La suddivisione può venir effettuata su una delle varie trascrizioni o direttamente sul segnale audio. In entrambi i casi deve essere eseguita sul flusso in lingua originale.

Questo task è presente in tutte le edizioni della campagna di valutazione tranne quella del 2004, in quanto il corpus di prova non comprendeva le fonti richieste.

★ Topic tracking task

Presente in ogni edizione della campagna, il task prevede l'associazione di una storia fornita al sistema ad un topic noto, detto topic target. Ogni topic target è definito nel sistema fornendo una o più storie che lo trattano.

Per supportare lo sviluppo di questo task vengono fornite, al fine di allenare il sistema, un insieme di storie la cui associazione ad un determinato topic è nota. Il task quindi si riduce a classificare in maniera corretta se le successive storie fornite in input siano pertinenti o no con quel topic.

Questo è l'unico task presente, rimasto praticamente invariato, in ognuna delle sette campagne di valutazione; da quella del 2000 in poi è classificato come primario.

★ Topic detection task

Presente dalla prima edizione fino a quella del 2003 è definito come il problema di identificare e successivamente riconoscere topic sconosciuti al sistema. Non c'è conoscenza a priori del topic da seguire ed il sistema deve avere coscienza di cosa sia, in generale, un topic.

Il sistema deve riconoscere un nuovo topic man mano che gli arrivano delle storie in input e, successivamente associargli eventuali altre storie "in topic" ricevute.

Questo processo costruisce un insieme di topic, definiti dalle associazioni tra storie.

★ First story detection task

Dal 1999 il NIST ha introdotto una nuova forma dell'on-line detection task presente nel progetto pilota del 1996. Il problema è quello di scoprire, all'interno di un flusso di storie provenienti da diverse fonti in più lingue ordinate cronologicamente, la prima storia che tratta un certo evento.

Questo task viene considerato come un sottoproblema del topic detection task; le differenze stanno principalmente nell'output prodotto.

Nella campagna di valutazione del 2004 questo task viene rinominato in "new event detection task".

★ Link detection task

Anch'esso presente dal 1999 e considerato task primario dal 2000 in poi, il link detection task consiste nel decidere se due storie parlano o meno di un certo topic. Come per il topic detection task il sistema deve avere una rappresentazione interna di cos'è un topic anche se, in effetti, non lo deve trattare esplicitamente.

I collegamenti creati tra le storie non sono considerati esclusivi, cioè una storia può, in linea di principio, parlare di più topic.

★ Supervised adaptive tracking task

Una delle novità introdotte nell'edizione del 2004 della valutazione è questa variante del topic detection task. In questo caso si rende disponibile ad un supervisore il valore di confidenza con la quale ha giudicato una storia "in topic" o "off topic", permettendo correzioni durante il tracking.

★ Hierarchical topic detection task

Forse la novità più eclatante della campagna del 2004 è stata la sostituzione del topic detection task con una sua versione gerarchica.

Questo approccio risolve due problemi riscontrati durante gli anni.

Il primo è che non tutti i topic hanno lo stesso livello di granularità, nel senso che uno può essere molto più generale di un altro e comprendere totalmente gli argomenti trattati da topic più specifici.

Il secondo problema nasce dall'assunzione che ogni storia tratta un solo argomento. Fino al 2004 infatti, i sistemi erano obbligati ad assegnare ogni storia ad uno ed un solo topic, creando quindi cluster di documenti con intersezione vuota e scartando le storie giudicate concernenti più topic.

La proposta è stata quella di una rappresentazione gerarchica in cui i cluster di documenti sono definiti a diversi livelli di granularità e le storie possono venir assegnate a più cluster contemporaneamente.

Le relazioni tra cluster sono espresse tramite un grafo diretto aciclico in cui la radice rappresenta il topic più generale di tutti, cioè l'intera collezione, e man mano che si scende si trovano nodi di sotto-topic più specifici, sottotopic del proprio padre, fino ad arrivare alle foglie che rappresentano i topic con granularità maggiore.

Questo nuovo modo di intendere eventi e topic è stato di fatto il motivo principale per cui è stato creato il corpus TDT5, seguendo nuove regole di annotazione.

Metodi di valutazione

Per misurare l'accuratezza degli algoritmi proposti dai ricercatori, sono stati elaborate, per ogni task, delle procedure atte a semplificare e rendere più chiari possibile gli obiettivi perseguiti dalle squadre partecipanti.

Ogni problema è formulato in termini statistici con la classica modalità per cui ad ogni decisione, oltre al responso, il sistema fornisce un punteggio che esprime la confidenza con la quale ha fornito la risposta.

I risultati sono presentati utilizzando le curve DET (detection error tradeoff, un esempio in figura 3), ottenute considerando due tipi di errori. Il primo è la *miss probability*, cioè il sistema ha classificato una storia "off topic" quando in verità era "in topic". Il secondo è la *false alarm probability*, cioè il sistema ha classificato una storia "in topic" quando in verità era "off topic".

Le probabilità sono ottenute dividendo il numero di volte in cui l'errore appare per il totale di storie analizzate.

Queste probabilità vengono misurate al variare di alcuni parametri di configurazione del sistema, come per esempio la soglia con la quale l'algoritmo valuta la similarità tra due storie.

Con le coppie di valori ottenute dai test è possibile tracciare dei grafici mettendo in ascissa la false alarm probability ed in ordinata la miss probability, ottenendo appunto le curve DET. Il risultato è molto chiaro: la curva più vicino all'origine è quella migliore.

Viene calcolato inoltre un costo combinando, secondo dei pesi, i due errori, dando, di solito, molta più importanza al primo. Intervengono inoltre le probabilità a priori di ogni storia di essere in topic oppure off topic. La cifra ottenuta viene normalizzata grazie ad un parametro ottenuto come il minimo punteggio tra quelli di un sistema che risponde sempre "si" ed uno che risponde sempre "no". Il primo per esempio avrebbe una miss probability dello 0%!

Grazie a questo costo è possibile determinare quindi la miglior combinazione miss/false alarm probability grazie alla quale si può risalire ai parametri ottimali con il quale sistema è stato tarato per eseguire i test.

Lo stato dell'arte

Verranno ora presentati alcuni significativi risultati scelti tra quelli ottenuti nell'edizione del 2004 della campagna di valutazione.

Hierarchical topic detection al TNO

Dei quattro sistemi proposti per il task della hierarchical topic detection il migliore è stato quello della Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek o, più brevemente, TNO.

Le tradizionali tecniche per la costruzione di cluster utilizzano una matrice delle distanze per realizzare una struttura di cluster ad albero binario. La complessità tipica è $O(n^2 log(n))$ in tempo e $O(n^2)$ in spazio, portando a circa 80Gb l'occupazione di memoria e

costringendo ad effettuare 80 miliardi di confronti tra documenti.

Queste ragioni hanno praticamente costretto a imboccare una strada diversa:

i. selezionare un campione del corpus;
ii. costruire un cluster gerarchico basato sul campione;
iii. ottimizzare l'albero ottenuto in modo da avere il costo minimo;
iv. assegnare i restanti documenti collegandoli alla struttura ottenuta.

i. Campionamento

Il primo passo è stato selezionare dal corpus 20000 documenti a caso, cifra che porta ad una matrice delle distanze di circa 800 Mb.

ii. Clustering

La tecnica di costruzione della struttura gerarchica si basa principalmente sulla cross-entropy, una misura strettamente legata alla divergenza informazionale, utilizzata per rappresentare i documenti. Viene generato un modello di riferimento, calcolato sull'intera collezione, grazie al quale vengono calcolate le misure di similarità utilizzate per riempire la matrice delle distanze. La struttura gerarchica è realizzata secondo la tecnica average linkage, cioè considerando la media di tutte le possibili distanze tra documenti appartenenti ai due cluster. Questa tecnica è risultata migliore di quelle del complete linkage o single linkage in quanto non soffre di effetti laterali come per esempio il fenomeno del concatenamento, per il quale al posto di un albero si ottiene una catena.

iii. Ottimizzazione

Il terzo passo consiste nell'ottimizzare la struttura gerarchica ottenuta, tipicamente un albero binario non bilanciato, secondo la metrica imposta dagli organizzatori.

Questa metrica consiste principalmente nel *travel cost* il quale valuta il percorso (inteso come numero di biforcazioni da imboccare e lunghezza dello stesso) da un generico cluster alla radice. Più basso è il risultato e migliore è la struttura visto che intuitivamente il generico fruitore che volesse ricevere informazioni su un determinato topic dovrebbe attraversare la gerarchia proprio lungo il percorso valutato.

Un ribilanciamento dell'albero rischia però di perdere informazioni sui topic stessi degradando il significato dei raggruppamenti.

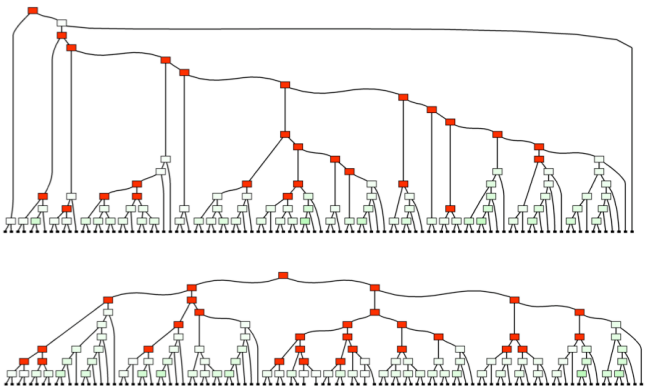

Figura 2: effetti del ribilanciamento su un cluster gerarchico. Le foglie rappresentano i documenti. I nodi rossi rappresentano: in alto quelli che verranno eliminati, in basso quelli creati per effetto dell'algoritmo.

La tecnica utilizzata è semplice: per prima cosa vengono rimossi i cluster i quali non hanno documenti direttamente connessi (quelli marchiati in rosso in figura 2) lasciando disconnessi un buon numero di cluster. Essi vengono utilizzati per costruire un nuovo albero ricorsivamente scegliendo i tre più piccoli ed unendoli, fino ad arrivare ad un unico cluster.

iv. Completamento della struttura

Viene costruito un indice dei documenti campione. I documenti del corpus non presenti nel campione vengono utilizzati come chiave per cercare all'interno di questo indice. Ai cluster che includono i 10 risultati migliori di questa ricerca viene associato il documento. Se un documento non produce risultati viene generato un nuovo cluster.

Supervised adaptive tracking alla CMU

La Carnegie Mellon University è una delle poche istituzioni che ha partecipato alla campagna di valutazione sin dalla prima edizione. Quando è stato proposto il nuovo task nel 2004 un loro gruppo ha subito accettato la sfida.

Il loro sistema utilizza un classificatore (cioè un sistema che assegna un documento ad una certa categoria, in questo caso "in topic" o "off topic") basato su due distinti algoritmi.

Inizialmente vengono scelti 30 documenti a caso per ogni topic marcati e si aspetta che il supervisore li segnali come "on topic" o meno.

Finchè la quantità di dati per far apprendere il sistema è basso viene usato il semplice classificatore di Rocchio che valuta semplicemente la vicinanza del documento al centroide degli esempi positivi e la lontananza dal centroide degli esempi negativi.

Quando la quantità di feedback del supervisore diventa sufficientemente alta il sistema cambia automaticamente classificatore utilizzandone uno a regressione logistica, più accurato del precedente sotto queste ipotesi.

Il lavoro svolto non è stato troppo faticoso: con poco sforzo hanno infatti adattato un sistema per l'*adaptive information filtering* presentato alla Text Retrieval Conference, riconoscendo che i due temi (*adaptive filtering* e *supervised tracking*) hanno moltissime cose in comune.

C'è da sottolineare inoltre che questo task ha messo in luce la netta superiorità, in termini di precisione, di un approccio supervisionato rispetto ad uno totalmente automatico.

Link detection alla UMASS

Alla University of Massachussets durante gli anni hanno proposto principalmente 2 approcci al link detection task, uno basato sul modello *vector-space* e l'altro sui *relevance model*.

Il primo utilizza una rappresentazione dei documenti secondo lo schema tf-idf, utilizzando cioè per ogni termine la sua frequenza all'interno del documento e il numero di documenti che lo includono, per poi misurare la similarità tramite le note tecniche del prodotto scalare, cioè del coseno.

Il secondo approccio utilizza un *relevance model* per stimare un modello linguistico da piccole porzioni di testo. Questo modello può essere visto come una specie di applicazione di tecniche di query-expansion, cioè trovare termini simili o vicini semanticamente o che comunque si possano trovare all'interno degli stessi documenti che comprendono i termini analizzati.

Per la stima del modello vengono scelti i 10 termini del documento che hanno la probabilità minore di essere pescati se si scegliessero 10 termini a caso nell'intera collezione. In qualche modo quindi si scelgono termini che compaiono quasi esclusivamente all'interno di quel documento.

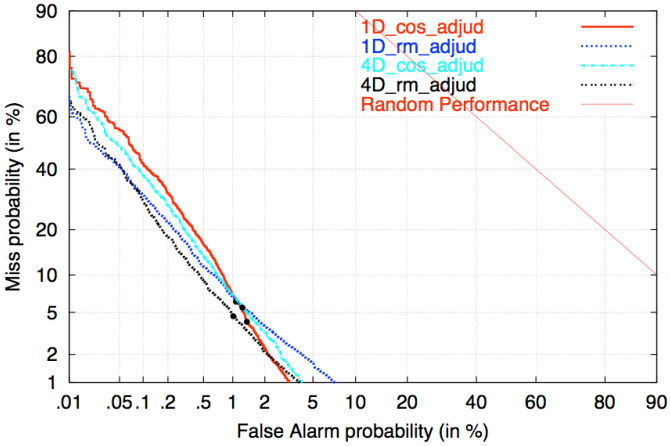

Figura 3: i quattro sistemi proposti da UMASS per TDT5

Una volta ottenute le due stime per i documenti in esame si misura la similarità tra i due tramite una complessa misura di divergenza simmetrica.

Quando i documenti sono entrambi nella stessa lingua non occorre ricorrere alle traduzioni automatiche, basta utilizzare i modelli statistici per la lingua in esame. Questa accortezza fa di questo sistema il migliore tra quelli mai presi in esame durante le sette campagne di valutazione.

Nella figura possiamo vedere le performance, in termini di curve DET, dei 4 sistemi proposti da UMASS alla settima campagna di valutazione in cui: 1D indica che i sistemi hanno utilizzato solo documenti tradotti automaticamente in inglese mentre 4D hanno eseguito i propri calcoli su quelli in lingua originale, quando possibile; cos indica l'utilizzo dello schema tf-idf mentre rm sta per relevance model.

Bibliografia

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. “Topic detection and tracking pilot study”, In Topic Detection and Tracking Workshop Report, 2001.

G. Doddington, "The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan," Available at http://www.nist.gov/speech/tests/tdt/index.htm, 1998

G. Doddington, “The 1999 Topic Detection and Tracking (TDT3) Task Denition and Evaluation Plan", Available at http://www.nist.gov/speech/tests/tdt/index.htm, 1999.

"The Year 2000 Topic Detection and Tracking (TDT2000) Task Definition and Evaluation Plan", Available at http://www.nist.gov/speech/tests/tdt/index.htm, 2000.

"The 2001 Topic Detection and Tracking (TDT2001) Task Definition and Evaluation Plan", Available at http://www.nist.gov/speech/tests/tdt/index.htm, 2001.

"The 2002 Topic Detection and Tracking (TDT2002) Task Definition and Evaluation Plan", Available at http://www.nist.gov/speech/tests/tdt/index.htm, 2002.

"The 2003 Topic Detection and Tracking (TDT2003) Task Definition and Evaluation Plan", Available at http://www.nist.gov/speech/tests/tdt/index.htm, 2003.

"The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan", Available at http://www.nist.gov/speech/tests/tdt/index.htm, 2004.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. “The DET curve in assessment of decision task performance”, In Proc. of ESCA 5th Eur. Conf. on Speech Comm. and Tech. - Euro-Speech '97, pages 1895-- 1898, 1997.

V. Lavrenko and W. Bruce Croft, “Relevance-based language models”, In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 120--127, 2001.

Y. Zhang, J. Callan, “CMU DIR Supervised Tracking Report”, 2004

D. Trieschnigg, W. Kraaij, “TNO Hierarchical topic detection report at TDT 2004”, 2004

M. Connell, A. Feng, G. Kumaran, H. Raghavan, C. Shah, J. Allan, “UMass at TDT 2004”, 2004

J. Allan, S. Harding, D. Fisher, A. Bolivar, S. Guzman-Lara, P. Amstutz, “Taking Topic Detection From Evaluation to Practice”, in Proceedings of the 38th Hawaii International Conference on System Sciences, 2005

S. Strassel, M. Glenn, J. Kong , “Creating the TDT5 Corpus and 2004 Evaluation Topics at LDC”, Available at www.ldc.upenn.edu/Projects/TDT5, 2004