WP3.1 Tecniche di analisi semantica per l'estrazione di ontologie bio-medicali

Autori: Eugenio Picchi
eugenio.picchi@ilc.cnr.it
Eva Sassolini
eva.sassolini@ilc.cnr.it
Sebastiana Cucurullo
nella.cucurullo@ilc.cnr.it
Monica Ensini
m.ensini@toscanalifesciences.org

Data: 20/06/2009 Versione: 1.0

Sommario

Descrizione generale

Il documento contiene la descrizione delle risorse e degli strumenti che Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC) ha messo a disposizione del progetto. L’obiettivo è quello di rendere accessibili alla regione Toscana, in particolare ai ricercatori e alle aziende toscane una serie di strumenti efficaci per lo sviluppo di nuove sinergie. Per la parte di nostra competenza abbiamo lavorato alla messa a punto di sistemi di analisi dei dati per una migliore estrazione delle informazioni; in questo senso abbiamo pensato a sistemi che non si limitassero a evidenziare cose o persone, ma che le mettessero in relazione tra loro.
Partendo dalla Base di Dati individuata nel WP1, cioè gli archivi delle *imprese* (intese come istituti pubblici o privati, enti di ricerca e società commerciali), delle *persone* (figure professionali legate alle imprese), gli archivi contenenti brevetti, pubblicazioni, studi clinici, accordi, etc., abbiamo estratto tutti i materiali testuali, collezionandoli in corpora di documenti finalizzati alla creazione delle risorse linguistiche.
I documenti così estratti sono stati sottoposti ad una serie di procedure di analisi automatica con lo scopo di identificare tutte le correlazioni più significative. Criteri di similitudine semantica hanno consentito poi di creare una rete di concetti in grado di fornire una navigazione più performante.
La creazione delle risorse linguistiche è funzionale allo sviluppo di strumenti di navigazione e di Information Retrieval in grado di sfruttarle, cioè a strumenti che catturano l’informazione, organizzandola, classificandola e distribuendola in modo funzionale agli obiettivi desiderati. Queste considerazioni hanno portato allo sviluppo di nuovi approcci basati su strumenti in grado di rappresentare, costruire e condividere la conoscenza, strumenti che portassero all’identificazione, l'annotazione e la classificazione di informazioni rilevanti nei testi e che rendessero possibile una classificazione semantica dei vari documenti analizzati secondo ontologie di settore.
La novità è rappresentata dal fatto che tutte le informazioni associate al testo vengono utilizzate per favorire una nuova modalità di navigazione, infatti il sistema si fa autonomamente carico di individuare eventuali caratterizzazioni semantiche e di suggerirle per un approfondimento della ricerca in ambiti sempre più definiti e dettagliati.

ILC per SUBITO

Gli strumenti messi a disposizione da ILC si possono ricondurre a due principali progetti, “Linguistic Miner” e “TextPower”, basati su moduli, metodologie e risorse sviluppate dall’istituto.

a) “Linguistic Miner” (LM) è nato per l’estrazione e l’acquisizione automatica di conoscenza linguistica da grandi collezioni di materiale testuale in lingua italiana con l’intento di sviluppare strumenti di analisi statistico-linguistica, procedure per la ricerca, il downloading e l’analisi automatica di grandi quantità di dati testuali dal WEB;

b) “TextPower” (TP), evoluzione del primo, ha l’obiettivo di individuare conoscenza semantica implicitamente presente nei documenti e di esplicitarla attraverso l’annotazione e la classificazione del testo. Sono parte fondamentale del conseguente arricchimento del testo le connotazioni terminologiche (mono-/polirematiche) e l’identificazione di “named entities” individuate con lo strumento di NER (Name Entity Recognition), anch’esso sviluppato nell’ambito di TP.

L'esperienza accumulata nel trattamento di grandi volumi di dati ha permesso, non solo il raffinamento degli strumenti di estrazione dell'informazione semanticamente rilevante, ma anche la creazione di risorse terminologiche corredo degli strumenti.
L'individuazione di tutte le componenti terminologiche di valore semantico ha contribuito alla realizzazione di un sistema di analisi, di classificazione e di navigazione nei testi, ma ha anche prodotto una rete di "conoscenza" in grado di essere utilizzata per funzioni di clustering e di navigazione "intelligente", funzioni che costituiscono la ricchezza dello strumento e del servizio offerto dal sistema.

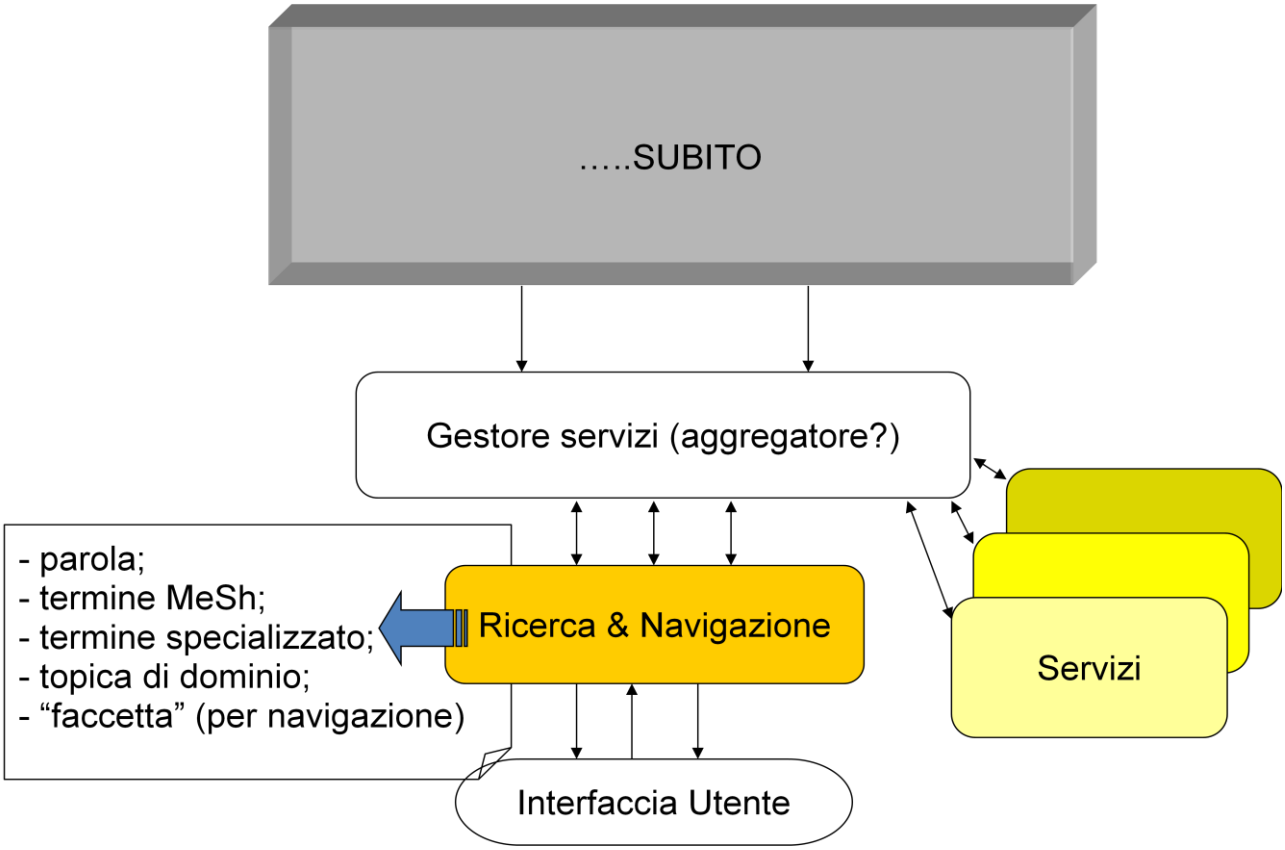

Figura 1: contributo ILC alle funzionalità del portale

Creazione delle risorse linguistiche

Creazione del corpus di riferimento

Nell'ambito del progetto "Linguistic Miner" sono stati sviluppati strumenti per l'analisi e la sintesi di evidenze linguistiche, puntando sulla convinzione che la conoscenza intesa come relazioni e dipendenze tra le informazioni contenute nei testi, possa essere estratta con l'aiuto di procedure statistiche. La caratteristica fondamentale del sistema è la capacità di utilizzare strumenti ed approcci linguistico-statistici per elaborare strategie di classificazione e riconoscimento di named entities e terminologia.

Analizzare il materiale testuale raccolto con l'aiuto di strumenti linguistici (morfologie e tagger) e di risorse linguistiche (dizionari terminologici, liste di nomi propri, cognomi, luoghi geografici,

istituzioni, ecc), è fondamentale per una produttiva applicazione delle funzioni statistiche di estrazione, che da sole non offrirebbero garanzie sufficienti per assicurare la bontà di quanto estratto. Il corpus necessario alla creazione delle risorse linguistiche è molto importante e costituisce la base di tutto il processo di creazione delle risorse specializzate e dell'adattamento delle procedure alle esigenze del progetto. Naturalmente non va dimenticata una fase redazionale finale che può anche suggerire nuovi adattamenti e modifiche all'intero processo. Secondo le specifiche del progetto sono basilari risorse linguistiche per la lingua inglese perché le più importanti risorse testuali del progetto (brevetti, pubblicazioni, studi clinici) sono reperibili in lingua inglese, ciò nonostante, si ritiene di strategica importanza per il progetto, la creazione di analoghe risorse per la lingua italiana, infatti non va dimenticato che il sito si pone l'obiettivo di creare sinergie tra ricercatori/aziende/persone in toscana.

per la lingua inglese

Per l'inglese i materiali testuali deputati alla formazione del corpus di riferimento sono stati estratti dalle fonti descritte nel WP1:

- Pubblicazioni PubMed;
- Brevetti internazionali (EP, US e WO);
- Studi Clinici (GOV, UE e IT).

In questo caso il training corpus è stato ricavato da 300.000 pubblicazioni estratte da PubMed, materiale che ci è stato messo a disposizione dagli altri partner del progetto. E' poi stato necessario un secondo corpus di testing per il controllo di quanto estratto e per l'eventuale riallineamento degli strumenti.

per la lingua italiana

Per l'italiano, lingua per la quale l'Istituto può mettere a disposizione un consolidato background di risorse e strumenti, non è stato possibile reperire gli stessi materiali che costituiscono la Base di Dati citata, in particolare non abbiamo trovato corrispettivi per l'archivio delle pubblicazioni (contenente tutti gli abstracts degli articoli presenti su PubMed) e per quello dei brevetti (informazioni brevettuali EP, US e Wo) che si trovano esclusivamente in inglese e forniscono il più cospicuo apporto testuale. E' stato quindi necessario reperire altro materiale testuale digitale di settore per creare il training corpus di riferimento. Nello specifico lo spider automatico è stato applicato a siti internet quali:

- Dipartimenti ospedalieri toscani;
- Dipartimenti universitari;
- Istituti del CNR;
- Sito del MIUR relativo alla lista dei progetti PRIN di settore approvati;
- Ecc..

Il corpus così formato è stato poi utilizzato per la fase di specializzazione delle risorse linguistiche. Per rispondere alle specifiche del progetto è stato necessario accrescere le risorse già patrimonio della linea di ricerca, in particolare è stata ampliata la componente più strettamente tecnico-scientifica della terminologia.

Estrazione di vocabolari di dominio

Per esigenze di uniformità con le risorse messe a disposizione dagli altri partner del progetto si è deciso unanimemente che la base di partenza per la creazione delle ontologie biomedicali fosse l'albero MeSh che, utilizzato anche per la categorizzazione dei documenti presenti in PubMed,

rappresenta un valido spunto di estrazione di conoscenza e di verifica delle associazioni semantiche estratte dai nostri strumenti statistico-linguistici.

Il Medical Subject Headings (MeSH) è un enorme vocabolario creato dal National Library of Medicine (NLM) degli Stati Uniti, con l'obiettivo di indicizzare la letteratura scientifica in ambito biomedico. MeSh è costituito da oltre 24.000 termini, organizzati gerarchicamente in sedici categorie identificate da lettere dell'alfabeto ("A" per l'anatomia, "B" per gli organismi, "C" per le malattie, ecc.[1]), ciascuna delle quali a sua volta è suddivisa in sottocategorie, identificate da numeri. Dalla struttura ramificata che scaturisce da questa classificazione, si ottengono alberi di termini. Per es., nel caso della categoria "C" (Malattie), avremo ventitré sottocategorie: "C01: Infezioni batteriche e micotiche", "C02: Malattie virali", "C03: Malattie parassitarie", "C04 Neoplasie", ecc., secondo lo schema seguente:

- Neoplasie [C04]
    - Neoplasie per sede Site [C04.588]
        - Neoplasie dell'apparato digerente [C04.588.274]
            - Neoplasie Gastrointestinali [C04.588.274.476]
                - Neoplasie dello stomaco [C04.588.274.476.767]

Tutti i nodi dell'albero, identificato come pertinente alle esigenze del progetto, sono stati utilizzati come base di partenza per la creazione di vocabolari di dominio. Per esempio nel caso delle malattie ("C") sono state scelte le sottocategorie principali e sono stati creati i 23 vocabolari corrispondenti a tutte le relative sottocategorie.

La presenza di questa terminologia ha permesso di individuare i materiali testuali candidati a diventare parte del training corpus per la specifica sottocategoria. Una volta ottenuto un corpus di riferimento è stato possibile individuare nei testi la terminologia che concorre alla formazione del vocabolario con l'attribuzione del corretto peso per ogni termine. Ogni vocabolario è stato ottenuto con procedure statistiche in grado di misurare l'attinenza di un termine al dominio ed è stato utilizzato per la creazione delle Topiche.

Terminologia

Per la classificazione del materiale testuale è necessaria l'estrazione di concetti, termini, istituzioni e nomi e ontologie di dominio. Il modulo semantico sviluppato nel task 3.2 utilizza tali ontologie per riconoscere, all'interno delle banche dati create in WP2/WP5, i documenti rilevanti. I dati annotati servono successivamente per creare le relazioni fra concetti (geni, proteine, tecnologie, malattie, cellule, farmaci, ecc.) e quindi rimandare al lettore/ricercatore/investitore un quadro completo delle competenze acquisite e disponibili.
I termini da estrarre non sono solo riconducibili a thesauri e vocabolari disponibili on-line ed appartenenti a diverse categorie come ad esempio geni, farmaci, molecole, ecc.. ma sono costituiti anche da quelli che le procedure di analisi semantica hanno recuperato.
Per esempio attraverso l'estrazione automatica di alberi ontologici:

– Acido (..acetico, ecc.)
– Agente (..immunosoppressore, ecc.)
– Alcol (..metilico, ecc.)

[1] Per le fasi del progetto che lo richiedevano abbiamo introdotto per semplicità alcuni esempi in italiano, naturalmente le strategie per la selezione dei nodi MeSh e le tecniche per l'estrazione del vocabolario di dominio sono state applicate anche ai materiali in lingua inglese

– Malattie rare

Oppure attraverso l'estrazione di termini (semplici e composti) riconducibili ad una terminologia di dominio:

agente immuno-soppressore, alterazione cromosomica

Infine ottenuti con l'estrazione di eventi:

aborto spontaneo, accrescimento tumorale, aggravamento della malattia, abbassamento della glicemia, collasso cardiocircolatorio

Creazione degli strumenti

Estrazione della terminologia

Nello specifico è stato necessario specializzare i moduli per l'acquisizione di tutta quella terminologia in grado di facilitare la navigazione nei materiali testuali, infatti, oltre all'acquisizione di tutti nodi MeSh presenti nell'albero, abbiamo corredato le risorse linguistiche di tutti quei termini che i vari moduli di estrazione d'informazione semanticamente rilevante hanno segnalato. Lo schema del processo di estrazione è sintetizzato nella figura sotto.

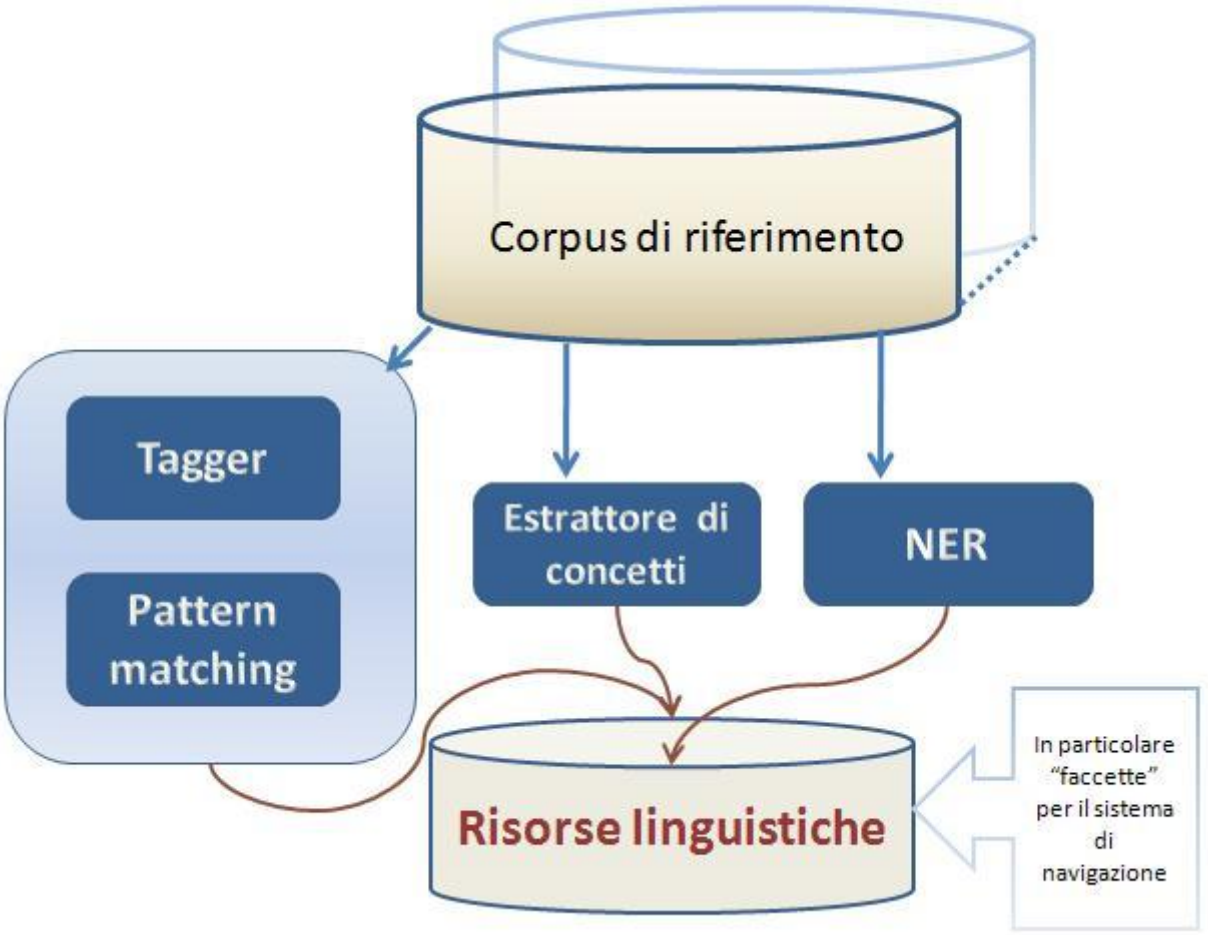

Figura 2: estrazione della terminologia di settore

Topiche

Estrazione di termini pivot partendo da MeSh

L'albero MeSH nella sua totalità contiene sottoalberi che, ad una prima analisi, sono stati considerati poco funzionali alla costruzione dei vocabolari di dominio. Questo ha portato alla decisione di acquisire tutti i nodi come terminologia ma di utilizzare solo alcuni per la creazione di Topiche in grado di filtrare uno specifico interesse dell'utente. Alla luce di queste considerazioni si è quindi proceduto ad operare una sfrondatura mirata con l'intenzione di selezionare le categorie MeSH di massima attinenza agli ambiti tecnologici e di ricerca nell'area biomedica. In termini pratici, di tutte le categorie MeSH si sono considerate le seguenti:

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Phenomena and Processes [G]
8. Disciplines and Occupations [H]

Le rimanenti categorie, listate di seguito, verranno vagliate e analizzate in un secondo momento, quando il progetto sarà nelle fasi più avanzate e sarà così possibile misurare le effettive esigenze dell'utente.

9. Anthropology, Education, Sociology and Social Phenomena [I]
10. Technology, Industry, Agriculture [J]
11. Humanities [K]
12. Information Science [L]
13. Named Groups [M]
14. Health Care [N]
15. Publication Characteristics [V]
16. Geographicals [Z]

In particolare, fra le otto categorie scelte, si è operata una ulteriore selezione all'interno della categoria "Disciplines and Occupations [H]" selezionando una sola delle due sottocategorie listate, vale a dire "Natural Science Disciplines [H01]. All'interno di quest'ultima ci si è per ora limitati ai nodi "Biological Science Disciplines [H01.158]", "Chemistry [H01.181]", "Microtechnology [H01.570]" e "Nanotechnology [H01.603]".

Creazione delle Topiche

Con il termine "Topica" identifichiamo un settore d'interesse scelto secondo le esigenze del progetto, questo è uno strumento che facilita la navigazione tra i materiali testuali, fornendo all'utente una selezione dei documenti pertinenti all'argomento cercato, infatti la procedura permette di esplicitare il dominio di appartenenza dei testi, sulla base del vocabolario terminologico identificato automaticamente e permette di misurare l'attinenza (ranking) di un articolo rispetto alla topica selezionata.
I moduli e le procedure che realizzano le Topiche si basano su tecniche linguistico-statistiche e sono finalizzati alla creazione di un vocabolario pesato che applicato su materiali testuali del progetto è in grado di produrre la selezione suddetta..

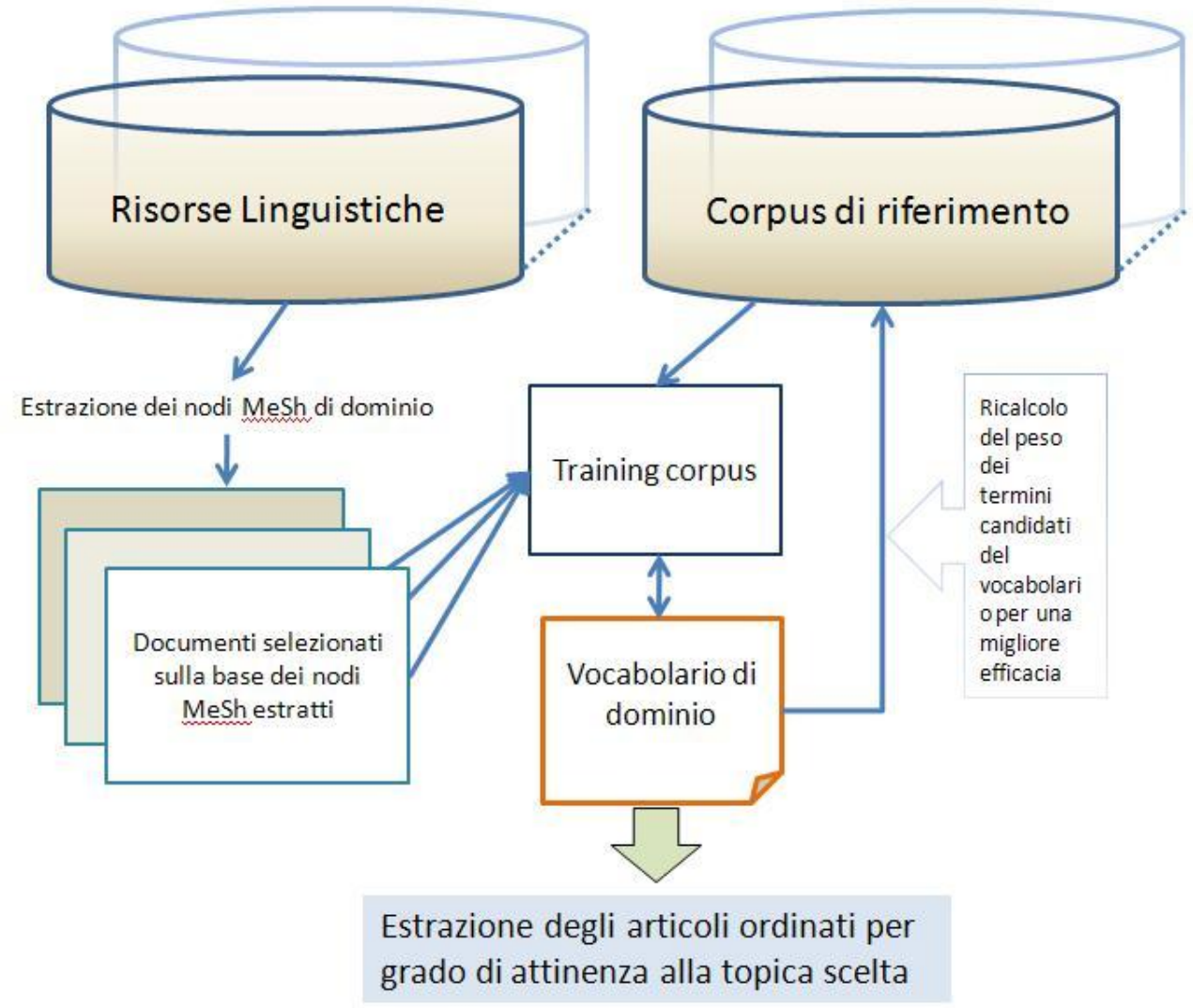

Figura 3: estrazione del vocabolario di dominio

Una volta creato il vocabolario, il sistema riconosce gli elementi costitutivi della Topica scelta, tra i materiali testuali disponibili e fornisce all'utente una lista ordinata per attinenza degli articoli ritenuti attinenti alla topica desiderata. Sostanzialmente è misurato il grado di fiducia con il quale si afferma che un articolo parla di quell'argomento. Varie fasi di testing hanno evidenziato un alto grado di successo della procedura, tuttavia la creazione di una Topica rimane una fase delicata e richiede una processo redazionale a posteriori: negli articoli troppo brevi, data l'esigua quantità di testo è possibile non pesare correttamente i termini, in relazione alle altre parole del testo. Questo richiede un'attenta valutazione dei casi e progressive fasi di tuning degli strumenti.

DBT & Faccette

La crescita del web, della tecnologia delle reti, della ricerca biomedica hanno dato un nuovo impulso allo sviluppo di applicazioni di Text Mining (TM), infatti attraverso l'uso di tecniche di Natural Language Processing (NLP) è possibile analizzare una grande quantità di testi. Per poter

trasformare l'informazione in conoscenza è necessario anzitutto assicurare un accesso flessibile e multi-dimensionale all'informazione stessa. Per questo motivo alla verticalità dei sistemi di catalogazione tradizionali e alla loro rigidità, la classificazione "a faccette" contrappone un sistema di classi (faccette) orizzontale e aperto.

La capacità di TextPower di arricchire il testo con forme esplicite di conoscenza, ha portato allo sviluppo di un sistema proprietario di categorizzazione dei testi "a faccette". Le faccette sono i termini riconosciuti ed esplicitati, come concetti semanticamente rilevanti in esso contenuti, e sono utilizzati per facilitare la navigazione nel testo. Questo permette di procedere ad un raffinamento della ricerca sulla base dei concetti che sono semanticamente correlati. Storicamente la tecnologia "a faccette" è stata utilizzata in biblioteconomia, per facilitare la navigazione nei cataloghi, in questo caso invece le faccette sono state ricavate utilizzando l'approccio, le risorse e le metodologie proprie di TP, nell'accezione terminologica che noi abbiamo utilizzato all'interno del nostro progetto "Text Power" e di conseguenza all'interno di SUBITO

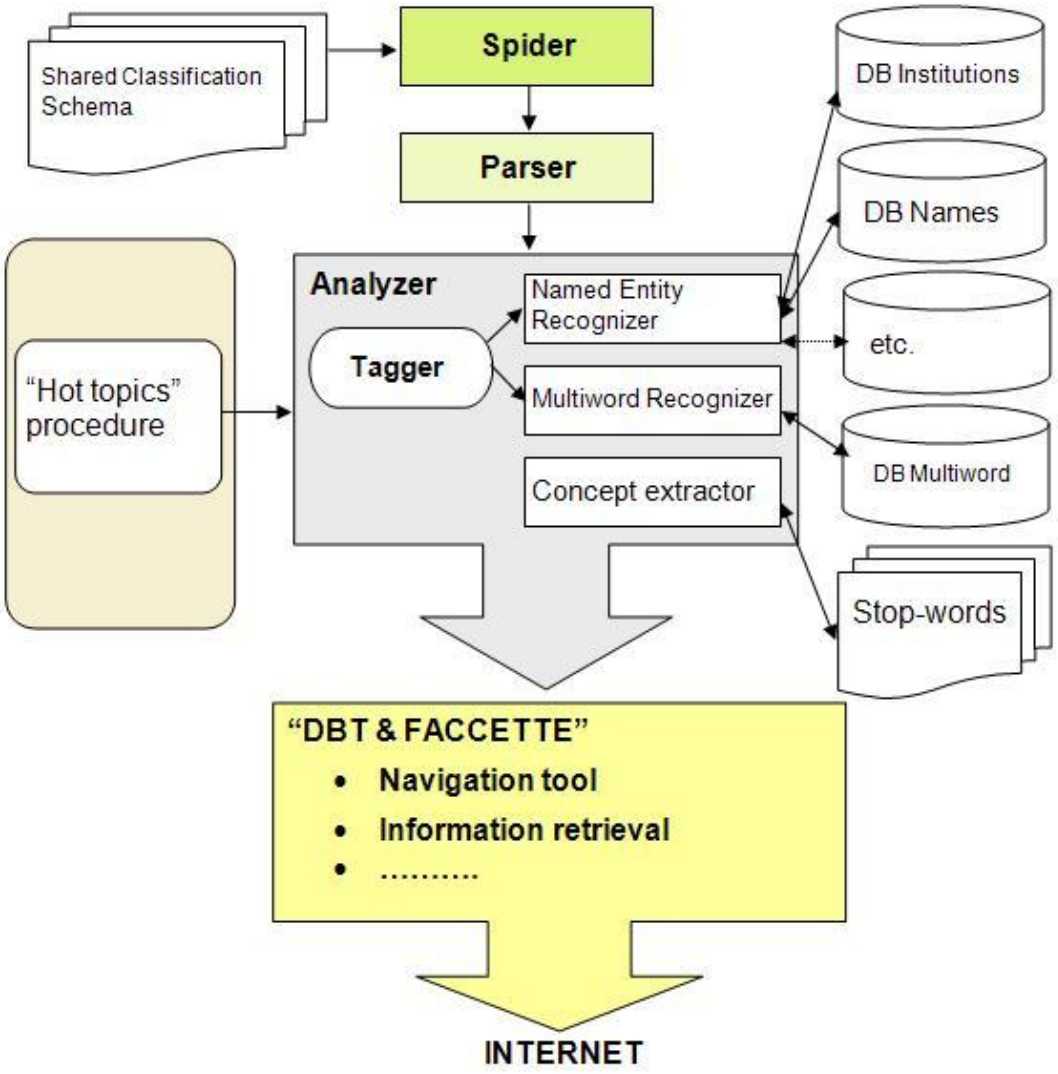

Figura 4: schema del progetto Text Power

Le caratteristiche principali di “DBT & Faccette” sono la categorizzazione dei testi “auto-adattiva”, che consente la ri-organizzazione automatica del contenuto sulla base dei concetti salienti e la capacità di esplicitare la tematica di appartenenza dei testi, sulla base del vocabolario terminologico identificato automaticamente. L’approccio permette all’utente di scoprire dinamicamente i concetti semanticamente rilevanti in un determinato dominio, e di procedere a raffinamenti della ricerca sulla base dei concetti semanticamente correlati sostanzialmente non esplicitati e virtualmente ignoti.

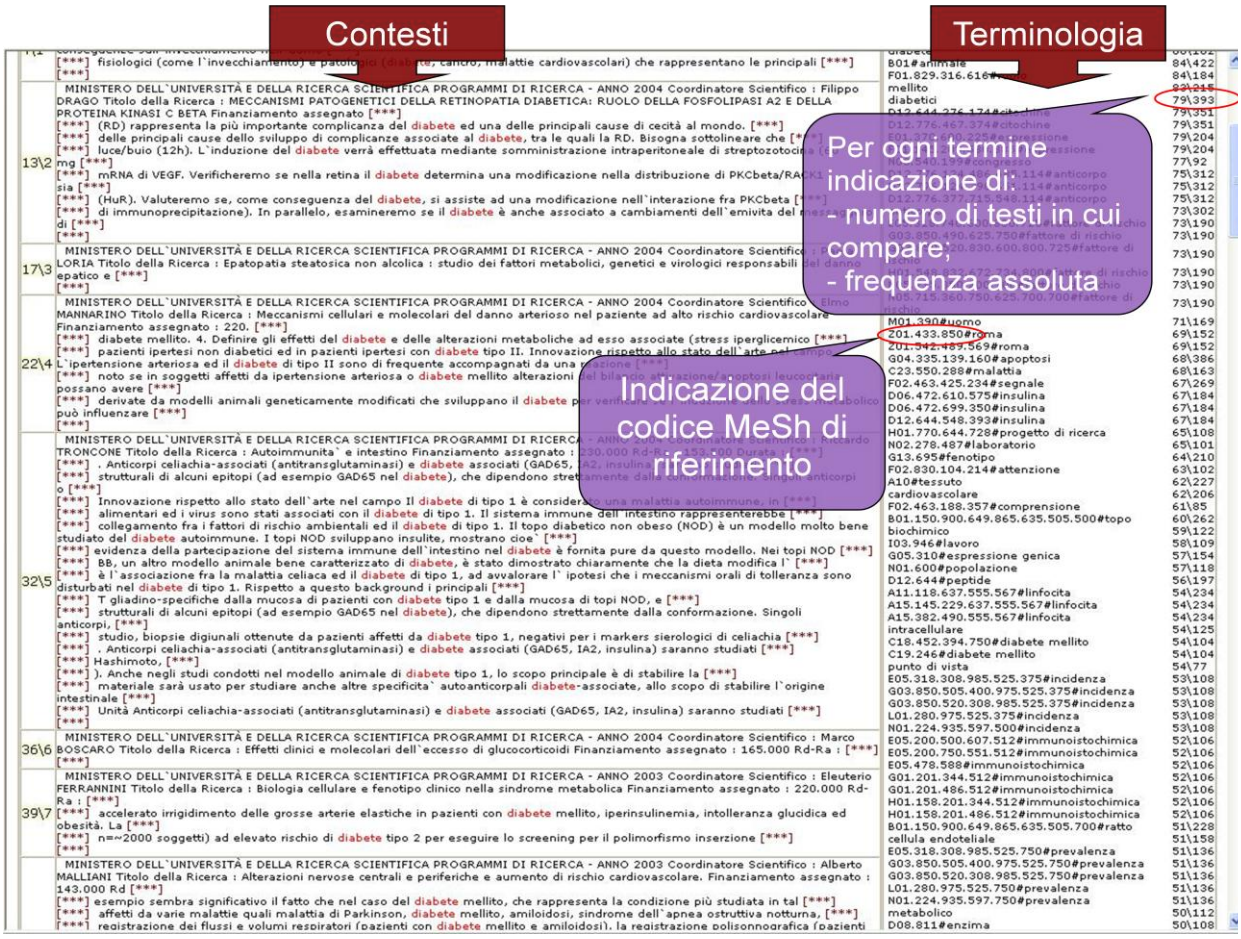

Figura 5: esempio di navigazione con le "faccette"