Costruzione di ontologie per sistemi di Information Extraction: un approccio terminologico

Fabio Massimo Zanzotto
Università di Roma "Tor Vergata",
Dipartimento di Informatica, Sistemi e Produzione,
00133 Roma,
zanzotto@info.uniroma2.it

Sommario L'estrazione dell'informazione (Information Extraction, IE) così come modellata nelle conferenze della serie Message Understanding Conference (MUC) [12, 13] è quell'attività di ridurre a *tuple sematiche* (i template) del materiale testuale appartenente ad un flusso informativo specifico. Seppure i sistemi sviluppati di IE nell'ambito delle MUCs si basano su metodologie diverse, la necessità di una profonda conoscenza dello scenario in cui questi sistemi devono operare è innegabile. Per superare il *collo di bottiglia conoscitivo* che limita l'applicazione dei sistemi di IE su larga scala occorrono sistemi che permettano di acquisire la conoscenza necessaria collezioni di testi riguardanti il dominio di interesse. Il limite maggiore di questi approcci esistenti è rappresentato dal fatto che necessitano della definizione a priori del bisogno informativo da soddisfare. Per superare questo limite, in questo articolo, proponiamo un approccio per la costruzione delle basi ontologiche per i sistemi di IE basato su tecniche di estrazione terminologica. L'applicazione di un modello di estrazione automatica di terminologia al particolare problema richiede un adattamento alle particolarità dei domini su cui i sistemi di IE sono generalemente applicati. Inoltre, la straodinaria importanza relazioni tra concetti tipiche di dominio impone la necessità di sviluppare tecniche automatiche per il loro riconoscimento ed estrazione. Viene dunque proposto un algoritmo innovativo per stimare l'importanza delle relazioni di dominio. Infine, viene presentato un caso di studio incentrato sulla produzione di una base di conoscenza ontologica per NAMIC, un sistema di accesso all'informazione testuale che propone un modello per la creazione di hyperlink tra notizie giornalistiche. Il sistema e l'analisi di ontologia prodotta vengono descritti.

1 Introduzione

L'estrazione dell'informazione (Information Extraction, IE) così come modellata nelle conferenze della serie Message Understanding Conference (MUC) [12, 13] è quell'attività di ridurre a *tuple sematiche* (i template) del materiale testuale appartenente ad un flusso informativo specifico. I template rappresentano il bisogno informativo che il sistema di IE deve soddisfare. Seppure i sistemi sviluppati di IE nell'ambito delle MUCs si basano su metodologie diverse (da una parte i sistemi cosidetti *profondi (deep)* dall'altra quelli *superficiali (shallow)* [9]), la necessità di una profonda conoscenza dello scenario in cui questi sistemi devono operare è innegabile.

Per superare il *collo di bottiglia conoscitivo* che limita l'applicazione dei sistemi di IE su larga scala occorrono sistemi che permettano di acquisire la conoscenza necessaria da depositi largamente disponibili dove questa sia naturalmente cristallizzata. Questi depositi sono naturalmente rappresentati da collezioni di testi riguardanti il dominio di interesse o, nel caso particolare della definizione di IE data nelle MUCs, riguardati lo scenario di applicazione.

Tutti gli approcci al problema tendono ad indurre da un insieme di documenti le regole di estrazione relative ad un particolare scenario. I documenti, se selezionati rispetto alla rilevanza per il particolare bisogno informativo (il template), rappresentano il modello di comportamento del linguaggio nel particolare dominio e, quindi, possono essere utilizzati per apprendere regole di estrazione dell'informazione saliente. Gli approcci di maggiore interesse sono quelli che prevedono il più piccolo grado di supervisione manuale dei documenti in ingresso poiché anche l'etichettamento manuale dell'informazione saliente all'interno dei documenti è un lavoro piuttosto costoso in termini di qualità di tempo che preziose risorse umane debbono spendere. In [14] si presentano dei risultati che misurano in 8 ore il costo temporale di etichettamento (con semplici strutture) di 160 piccoli documenti. In [14, 17] vengono presentati dei modelli che si occupano dell'estrazione di pattern più o meno complessi nell'ambito di un classico sistema di IE applicato ad un flusso di documenti relativi ad un particolare scenario. I documenti di interesse sono generalmente considerati come contenti un solo evento. In ambedue gli approcci si trae vantaggio dal concetto di importanza di un documento rispetto al bisogno informativo considerando l'insieme dei documenti come partizionato [14] o patizionabile [17] in documenti rilevanti e documenti non rilevanti rispetto a questo bisogno.

Il limite maggiore di questi approcci è che necessitano della definizione a priori del bisogno informativo da soddisfare. Questo limita l'applicabilità dei sistemi di apprendimento e dei risultanti sistemi di IE a scenari particolari e noti.

Per superare questo limite, in questo articolo, proponiamo un approccio per il problema della costruzione delle basi ontologiche per i sistemi di IE basato su tecniche di estrazione terminologica. L'applicazione di un approccio di estrazione di terminologia al problema specifico favorisce un cambiamento di attitudine nell'estrazione di pattern sintattico-lessicali. Infatti, nell'estrazione di terminologia per un dominio, la sorgente principale di nuova informazione è il modello esteso del dominio stesso. La stessa collezione coerente di testi fornisce infatti indicazioni sui bisogni informativi che può soddisfare che sono poi i concetti tipici del dominio stesso. Tuttavia, l'applicazione di un modello di estrazione automatica di terminologia al particolare problema richiede un adattamento come verrà discusso nella Sez. 2. Infatti, le particolarità dei domini su cui i sistemi di IE sono generalemente applicati richiederà alcuni aggiustamenti nel modello di estrazione. Inoltre, la straodinaria importanza relazioni tra concetti tipiche di dominio impone la necessità di sviluppare tecniche automatiche per il loro riconoscimento ed estrazione. Come verrà discusso nella sez. 2.3, il problema può essere scomposto come è stato fatto per il riconoscimento dei termini: da una parte occorre definire la forma superficiale ammissibile e dall'altra una funzione di stima dell'importanza per il dominio. Un algoritmo innovativo per stimare l'importanza delle relazioni di dominio viene proposto nella Sez. 3.

Infine, viene presentato un caso di studio incentrato sulla produzione di una base di conoscenza ontologica per NAMIC, un sistema di accesso all'informazione testuale che propone un modello per la creazione di hyperlink tra notizie giornalistiche. Il sistema e l'analisi di ontologia prodotta vengono descritti nella Sez. 4.

2 Un approccio terminologico all'acquisizione di pattern per l'IE

Volendo quindi spingere sull'ampiamento del campo di azione dei sistemi di IE, anche in vista dell'applicazione a problemi meno strutturati come il Question Answering e la Document Summarization per i quali il dominio non è determinato a priori, proponiamo un approccio alla costruzione di ontologie di IE che sia il più possibile non supervisionato basato sull'applicazione dei concetti e delle metodologie di estrazione sviluppate per l'analisi terminologica dei domini. L'applicazione di un approccio di estrazione di terminologia al problema della costruzione di una ontologia applicabile per il compito di IE impone un cambiamento di attitudine nell'estrazione di pattern sintattico-lessicali come quelli descritti in [14, 17]. Infatti, nell'estrazione di terminologia per un dominio, la sorgente principale di nuova informazione è il *modello esteso del dominio* stesso. Questo modello esteso che è composto da una collezione di documenti relativi all'argomento in questione, una terminologia pre-esistente all'analisi ed, eventualmente, una terminologia ed una non-terminologia prodotta durante l'analisi, è il punto di partenza da cui gli strumenti automatici così come i terminologi possono indurre una conoscenza intenzionale del dominio. Al contrario, il punto di partenza per la costruzione dei sistemi di IE è un *bisogno informativo* particolare, conosciuto e generalmente molto ristretto come, ad esempio, il *lancio di missili spaziali* di una delle MUC. Questi bisogni informativi particolari risultano influenzare la costruzione delle basi ontologiche sottostanti ai sistemi di IE e, conseguentemente, dei meccanismi di apprendimento di queste ultime. In particolare, gli algoritmi di apprendimento possono beneficiare di una classificazione dei documenti da cui apprendere che sia basata sull'importanza degli stessi rispetto al particolare e ristretto bisogno informativo. Questa ipotesi di rilevanza può essere poi efficacemente sfruttata per indurre dai documenti le proprietà testuali osservabili che possono contribuire a costruire la base di conoscenza [16, 15]. Purtroppo, generalmente le metodologie di apprendimento, pur essendo adattabili a differenti bisogni informativi, sono sviluppate con in mente il fatto che il bisogno informativo da soddisfare è un particolare tipo di evento in un dominio conoscitivo molto più grande (per esempio, i cambiamenti del top management di aziende in uno dominio conoscitivo finanziario).

Le metodologie di apprendimento rispecchiano la filosofia (e le specifiche) imposte al task di IE nelle conferenze MUCs ovvero, come detto già in precedenza, dato un bisogno informativo e il template relativo, costruire un sistema che possa soddisfarlo. Al contrario, lo scopo della modellazione del dominio conoscitivo fatto nell'ambito dell'analisi terminologica è quello di estrarre, dato un modello esteso dello stesso dominio, una forma distillata di conoscenza sotto forma di un gran numero di concetti e di relazioni tra di essi (un tesaurus, un dizionario o una ontologia di dominio). Data la collezione di documenti, la conoscenza in questi racchiusa e, dunque, i bisogni informativi che questi possono soddisfare non sono chiari ma debbono essere scoperti durante l'analisi. Questa analisi dei documenti non è guidata verso uno scopo come quello di riempire un template prestabilito ma è orientata allo scoprire la conoscenza contenuta nell'insieme dei documenti. Quindi, è l'analisi stessa dei documenti che chiarifica in una forma sintetica l'insieme di concetti che possono essere individuati e cercati all'interno della base testuale.

Proponendo una visione incentrata sul corpus piuttosto che sul bisogno informativo, il punto di osservazione "terminologico" apre dunque prospettive diverse sul come l'IE può essere inteso e modellato. In particolare, per quanto riguarda il problema dell'apprendimento della conoscenza, induce a pensare algoritmi che siano aperti a domini conoscitivi sconosciuti a priori e ad un apprendimento su larga scala. Pertanto questa è una strada percorribile anche

per i "nuovi" paradigmi di estrazione dell'informazione (come il Question Answering o la General Purpose Document Summarization).

L'applicazione della visione terminologica alla ricerca automatica di pattern per l'IE richiede in ogni caso che siano definiti quali siano i concetti e le relazioni tra concetti che debbono essere modellati all'interno della base ontologica di dominio e, in conseguenza, quali siano le proprietà osservabili all'interno dei testi che gli algoritmi di ricerca debbono selezionare e far emergere. Nel paragrafo successivo (Sez. 2.1) discuteremo le differenze principali che esistono tra le basi di conoscenza terminologiche "classiche" (TKB) e ciò che è invece atteso come dover essere presente nelle ontologie per sistemi di IE. Nelle Sez. 2.2 e Sez. 2.3 vengono invece discusse più specificamente i concetti e le relazioni tra concetti di interesse. Successivamente, nel paragrafo 3, viene descritta una metodologia innovativa ed efficiente per l'estrazione di relazioni dal modello esteso del dominio.

2.1 Le ontologie per l'IE rispetto alle "classiche" TKB

La differenza di contenuto delle basi di conoscenza per i sistemi di IE rispetto a quello tipico delle basi di conoscenza terminologiche dipende dalla differente prospettiva per cui queste stesse basi di conoscenza vengono prodotte. Da una parte le ontologie per l'IE vengono prodotte per modellare domini giornalistici al fine di scoprire un fatto che è una risposta alle classiche domande 5W mentre le seconde sono state "storicamente" sviluppate con l'ottica di modellare domini scientifici o tecnici. Di conseguenza, i concetti e le relazioni e, dunque, le proprietà superficiali osservabili di interesse per le due discipline differiscono. Nel dominio giornalistico e con l'obbiettivo di rappresentare un fatto accaduto, gli *attori dei fatti* sono generalmente entità uniche denotate da un nome (named entities). Questi ultimi sono referenziati nei testi attraverso dei nomi propri. Nell'istanza del fatto, l'entità coinvolta è importante (come ad esempio la *Acme Inc.*) ma nella descrizione della regola per l'estrazione dell'informazione di maggiore importanza è la classe di appartenenza dell'entità stessa ovvero è più importante stabilire che esiste il concetto di `COMPANY's shares` rispetto al verificare l'esistenza del concetto `Acme Inc.'s shares`. L'interesse per la forma superficiale è dunque più basso rispetto a quello che viene riservato per la scoperta di concetti tecnici o scientifici nell'ambito dei classici domini dell'estrazione della terminologia. Si noti che anche nel caso in cui siano coinvolte entità nominate in questi ambiti non è importante la classe di appartenenza dell'entità nominata ma è importante l'istanza. Si prenda come esempio le leggi scientifiche come: *Zipf's law*, *Newtown's law*, etc. Inoltre, in un ambiente di IE, le relazioni tra i concetti tipiche del dominio sono quelle di maggiore interesse siano esse rappresentate in forme superficiali espresse in sintagmi nominali che in sintagmi verbali. In particolare, sono le relazioni tra concetti che permettono di individuare i concetti e le entità nominali che vengono poi utilizzate per riempire il template. Gli aspetti e i modelli delle forme di espressione delle relazioni di dominio tra concetti giocano un ruolo imporante per i modelli di estrazione delle basi di conoscenza ontologiche che soggiacciono ai sistemi di IE.

Anche se la differenza sostanziale mostrata sembrerebbe rendere inapplicabile il modello di estrazione di terminologia mostrato nei capitoli precedenti al problema della costruzione di ontologie per i sistemi di IE, nelle prossime sezioni mostreremo come sia possibile adattare con piccoli accorgimenti i concetti di *forma superficiale di interesse* e *importanza per il dominio* al problema specifico. Riteniamo infatti che l'adozione di una filosofia di estrazione terminologica centrata sull'utilizzo di un modello esteso di dominio permetta di cambiare l'ot-

tica di approcci all'IE: i bisogni informativi soddisfacibili sono quelli le cui risposte possono emergere dai documenti a disposizione. In particolare, tecniche di estrazione terminologica possono essere utilizzate per costruire ontologie di dominio su larga scala.

La possibilità di applicazione delle tecniche sviluppate per l'estrazione di terminologia richiede per poter dunque essere applicata che siano chiarificate quali siano le forme superficiali e le rappresentazioni interne dei concetti da modellare e quali invece siano le forme e le rappresentazioni relative alle relazioni tra i concetti tipiche di dominio che vogliono essere modellate.

2.2 I concetti: i nodi della gerarchia

I domini conoscitivi in cui si muove un sistema di IE richiedono, come abbiamo già evidenziato, di conoscere differenti tipi di concetti. In particolare il concetto rappresentato all'interno della base di conoscenza può essere una sussunzione dei concetti espressi nel modello esteso del dominio la cui specializzazione è interessante solo all'atto di utilizzo della conoscenza ontologica per l'estrazione effettiva di informazione. Infatti, in un dominio finanziario, è di interesse sapere che esiste il concetto di `COMPANY's shares` ma all'atto del recupero dell'informazione è interessante sapere che in quel testo è di *Acme Inc.'s shares* che si parla e che sono un particolare tipo di `COMPANY's shares`. In particolare, questi concetti di cui è interessante conoscere una generalizzazione coinvolgono da entità dette nominate (*named entity*). Queste ultime sono raggruppate in classi di interesse che, purtroppo, variano da un dominio conoscitivo all'altro. In tutti i domini coinvolgono comunque quelle entità rappresentate nel testo tramite l'utilizzo di nomi propri e l'utilizzo si numeri.

In una ontologia per un sistema di IE, i concetti di interesse sono dunque di tre tipologie:

- le classi di appartenenza delle entità nominate;
- i classici concetti terminologici come quelli veicolati da forme superficiali come i sintagmi nominali, ad esempio `pre-tax profit`, `profit warning`, oppure `intermin dividend`;
- i concetti terminologici estesi che integrano un concetto terminologico ed una classe di appartenenza di entità nominate come ad esempio `pre-tax profit of CURRENCY` o `COMPANY's shares` dove `CURRENCY` è una quantità di soldi.

Si noti che esiste un insieme possibile di default di classi di appartenenza delle entità nominate ovvero quello composto da due sole classi: `NAMED-ENTITY` e `NUMBER`. Questa osservazione permette di applicare il metodo di estrazione anche a domini per i quali non sia definito a priori questo insieme.

Se l'interesse è dunque su questa tipologia di concetti, occorre verificare che sia possibile farli emergere dal modello esteso del dominio attraverso gli strumenti costruiti per l'analisi terminologica. Fissando dunque il fatto che la funzione di importanza sia o la frequenza di apparizione della forma superficiale di un determinato concetto o la frequenza contrastiva [2], occorre stabilire se il meccanismo per recuperare le forme superficiali sia applicabile e se esiste un metodo per estenderlo al fine di permettere di esprimere il legame tra la forma superficiale occorsa e il concetto generalizzato che si vuole modellare nella base di conoscenza. Si noti che si desidera valutare l'importanza del concetto generalizzato che si vuole mettere

nella base di conoscenza e non l'importanza della particolare apparizione. Inoltre, le classi di appartenenza delle entità nominate sono di diritto importanti per il dominio.

Il modulo estrazione di forme superficiali descritto in [2] continua ad essere un valido terminatore della catena di parsing come evidenziatore di forme superficiali di concetti potenzialmente interessanti a patto che i vincoli di selezione delle forme superficiali siano adattate al particolare ambiente. Occorre solamente descrivere il meccanismo di correlazione del concetto espresso dalla forma superficiale al concetto generale che si vuole esprimere nella base di conoscenza. Se appare *pre-tax profit of 23 million dollars* lo si vuole relazionare con `pre-tax profit of CURRENCY`. Il modulo di estrazione permette di identificare sintagmi parziali $p = (CS_p, l_p)$ che soddisfino dei vincoli. Dato dunque p è immediato trovare il concetto più generico al quale è relazionato che può essere denotato dalla sequenza:

$$d(p) = s_1 s_2 ... s_n \tag{1}$$

con s_i forma canonica di $C_i \in CS_p$, ovvero $s_i = fc(C_i)$. La forma canonica è così ottenuta:

$$fc(C_i) = \begin{cases} fs(C_i) & \text{se } C_i \in CS \\ classNE(C_i) & \text{se } C_i \in NE \\ fc(C_i^1)...fc(C_i^m) & \text{se } C_i \in CC \text{ e } sons(C_i) = C_i^1, ..., C_i^m \end{cases} \tag{2}$$

dove CS e CC sono rispettivamente la classe dei costituenti semplici e quella dei complessi, $fs(C_i)$ è la forma superficiale di C_i, $classNE(C_i)$ è la classe dell'entità nominata contenuta in C_i e $sons(C_i)$ sono i sottocostituenti del costituente complesso C_i. Per la testa $h(p)$ di p, la $fc(h(p))$ seleziona solo le parti che soddisfano i vincoli.

Sia la funzione di importanza basata solo sulla frequenza che quella basata sull'analisi contrastiva non richiedono particolari algoritmi per essere contate efficientemente. L'insieme dei concetti reputati importanti per il dominio viene chiamato $Conc$.

2.3 Relazioni di dominio tra concetti

Rispetto all'estrazione delle relazioni generali tra concetti del dominio tipo `ISA` e `PART-OF` [11, 6, 7], in questo paragrafo si vogliono studiare e proporre dei metodi per estrarre le relazioni che non siano generiche ma siano tipiche di dominio in analisi. Pertanto, è necessario sia stabilire quali siano le forme prototipiche ammissibili con cui le relazioni si manifestano nel modello esteso del dominio sia quale sia il metro per stabilire l'importanza di tali relazioni rispetto al dominio.

Le relazioni di cui si vuole studiare il comportamento e che si vogliono individuare nel corpus di dominio sono quelle di tipo funzionale cioé relazioni n-arie che legano concetti tipici del dominio. Ad esempio, una relazione del tipo

Es. 3. *Una relazione di* `sell`*.*
```
sell(someone,something,(to,company),(for,amount_of_money))
```

può essere di interesse in un campo finanziario in un ambito di estrazione di informazione da un flusso di notizie. Tuttavia, così come i termini possono avere delle varianti, una relazione del genere può manifestarsi in differenti modi nell'ambiente testuale:

- l'ordine degli argomenti della relazione può non essere fissato;
- la relazione può essere espressa da differenti tipologie di frasi (frasi nominali o frasi verbali);
- la testa della frase che rappresenta la relazione può variare.

In questo lavoro analizziamo il primo problema assumendo che le frasi di interesse siano quelle verbali e si considera come un problema successivo quello di stabilire l'equivalenza tra le frasi verbali rette da teste differenti.

Se da una parte è risolto il problema di stabilire che le forme superficiali di interesse delle relazioni specifiche sono quelle rappresentate da sintagmi retti da verbi, dall'altra parte rimane aperto quello di definire una funzione di stima dell'importanza per il dominio per queste relazioni. Questa funzione si deve basare in ogni caso su una forma di calcolo della frequenza anche se fosse orientata alla stima della differenza tra domini come quella presentata in [2]. Per stabilire dunque l'importanza di una relazione del tipo in Es. 3 occorre in prima istanza valutare quante volte è apparsa nel corpus del modello esteso per poi comparare questo valore con quello delle altre relazioni che compaino nel corpus. Dunque, sia in fase di estrazione delle relazioni specifiche di dominio dal corpus che in fase di applicazione di queste per l'estrazione di legami tra concetti occorre stabilire l'equalinza tra differenti istanze della relazione stessa. Nel caso della relazione nell'es. 3 occorre stabilire che le seguenti frasi sono una sua istanza:

Es. 4. *Financial News*

(a) Eon, the German utility formed by the merger of Veba and Viag, is poised to sell its electronics arm to an Anglo-American consortium for about $2.3bn.

(b) It is understood to be near a deal to sell the Longview smelter for $150m to McCook Metals.

Si può incominciare a notare che la libertà di movimento degli argomenti verbali (come `for CURRENCY` o `to COMPANY`) e la loro eventuale mancanza possa generare problemi computazionalmente non risolvibili facilmente. All'estrazione di queste particolari forme viene dedicata la sezione successiva.

3 Estrazione di relazioni di dominio: forma normale e funzione di importanza

La ricerca delle relazioni importanti per il dominio purtroppo non è così semplice perchè lo spazio delle relazioni possibili da analizzare è piuttosto grande. Non si tratta infatti di computare, dato un insieme R di relazioni stabilite, quale sia la frequenza di ciascuna di esse nel corpus C al fine di stabilire quale siano le più importanti in quel determinato dominio. Al contrario, dato un corpus C occorre costruire valutare le possibili relazioni che si instaurano tra i concetti ed estrarre quelle che risultano essere più interessanti ovvero quelle che riescono a superare una certa soglia di frequenza. Data quindi una rappresentazione sistematica degli eventi che capitano nel corpus bisogna fornire delle procedure efficienti che permettano di

contare gli eventi apparsi con una complessità computazionale bassa e dunque che permettano di ottenere i risultati in un tempo ragionevole.

Rispetto a quanto fatto nell'analisi dei concetti espressi da sintagmi nominali, la computazione della frequenza è più complessa. Infatti, mentre nel caso dei sintagmi nominali si tratta di analizzare la frequenza di n-grammi di parole supposti continui, nell'analisi delle relazioni occorre stabilire la frequenza di n-grammi di parole supposti discontinui. In questo secondo caso, lo spazio di analisi risulta essere decisamente più grande. Più in particolare in questo spazio non ci si può nemmeno affidare alla obliquità degli argomenti verbali. Infatti, la soluzione adottata in [16] che prevedendo una obliquità fissata per gli argomenti verbali (cioè soggetto, oggetto) permette una loro agevole conta non tiene in considerazione due fenomeni: da una parte, la naturale libertà di movimento degli argomenti verbali presente sia in lingue come l'inglese ma più evidente in lingue romanze come l'italiano e, dall'altra, l'impossibilità di avere parser sintattici che stabiliscano con certezza i modificatori verbali e quelli invece che non lo sono. Un esempio di movimento degli argomenti verbali è dato in es. 4 dove gli argomenti retti dal *to* e dal *for* del verbo *to sell* appaiono in differente ordine.

3.1 Lo spazio delle istanze delle relazioni: forma normale

Data dunque la rappresentazione C del corpus in cui le istanze delle relazioni siano normalizzate, cioè siano viste attraverso coppie $(v, (a_1, ..., a_n))$ dove v è il verbo che regge il sintagma e $(a_1, ..., a_n)$ son gli n argomenti verbali, si vuole studiare quale siano le relazioni $(r, \{ra_1, ..., ra_m\})$ che in questo corpus si sono manifestate e quale sia la loro frequenza di apparizione. E' importante notare che, mentre l'ordine degli argomenti delle forme normali delle istanze delle relazioni è importante, l'ordine degli argomenti della relazione è artificioso e, in fase di rappresentazione finale lineare, indifferente. Definendo dunque $C(v)$ come la sottoparte del corpus che si riferisce al verbo v, ovvero:

$$C(v) = \{(a_1, ..., a_n) | (v, (a_1, ..., a_n)) \in C\} \tag{5}$$

e come $A_\Lambda(v)$ e $A_\Sigma(v)$ rispettivamente gli argomenti lessicalizzati e gli argomenti puramente sintattici che il verbo v manifesta nel corpus $C(v)$, ovvero:

$$A_\Lambda(v) = \{a | \exists (a_1, ..., a_n) \in C(v) \wedge \exists i.a_i = a\} \tag{6}$$

$$A_\Sigma(v) = \{s | \exists ((s_1, l_1), ..., (s_n, l_n)) \in C(v) \wedge \exists i.s_i = s\} \tag{7}$$

quello che si vuole studiare è la frequenza delle relazioni $R(v)$ rette dal verbo v. Le relazioni $R(v)$ risultano essere definite come un insieme di argomenti lessicalizzati o sintattici di cardinalità al massimo $MC(v)$, cioè il massimo numero di argomenti che il verbo v manifesta negli esempi contenuti nel corpus di interesse. Formalmente, l'insieme delle relazioni $R(v)$ del verbo è l'insieme seguente:

$$R(v) = \bigcup_{i=1...MC(v)} R_i(v) \tag{8}$$

dove $R_i(v)$ è l'insieme di tutte le combinazioni con ripetizione di i oggetti elementi dell'insieme $A(v) = A_\Lambda(v) \cup A_\Sigma(v)$. Data questa definizione, ciò che si vuole stabilire è la frequenza nel corpus $C(v)$ di ogni $r(v) \in R(v)$. Il contesto $c = ((s_1, l_1), ..., (s_n, l_n))$ viene considerato essere una istanza della relazione $r = \{ra_1, ..., ra_m\}$ di interesse se gli argomenti di r appaiono in un qualsiasi ordine come argomenti di c ovvero se per ogni i esiste un j tale che $(s_i, l_i) = a_i$ se $r_i \in A_\Lambda(v)$ oppure $s_i = a_i$ se $r_i \in A_\Sigma(v)$. Prendendo ad esempio come corpus le frasi dell'Es. 4 e il verbo *sell* come obbiettivo dell'analisi, il corpus $C(sell)$ è composto dai due seguenti contesti normalizzati:

$$C(sell) = \{ \begin{array}{l} ((dirobj, arm), (to, companyNE), (for, currencyNE)), \\ ((dirobj, smelter), (for, currencyNE), (to, companyNE))\} \end{array} \quad (9)$$

mentre gli argomenti possibili di *sell* sono:

$$A(sell) = \{ \begin{array}{l} (dirobj, arm), (dirobj, smelter), (to, companyNE), \\ (for, currencyNE)), dirobj, to, for\} \end{array} \quad (10)$$

Dal momento che il contesto di estensione massima è composto in questo da 3 argomenti, le relazioni da esaminare sono quelle con al massimo 3 argomenti che variano nell'insieme $A(sell)$. Ciò che si vuole ottenere infine è una conta delle apparizioni delle varie relazioni come mostrato nella tabella seguente:

Relazione	**Frequenza**
$((to, companyNE))$	2
$((for, currencyNE))$	2
$(dirobj, (to, companyNE), (for, currencyNE))$	2
$((to, companyNE), (for, currencyNE))$	2
$((dirobj, arm), (to, companyNE), (for, currencyNE))$	1
$((dirobj, smelter), (to, companyNE), (for, currencyNE))$	1

La normalizzazione dell'insieme dei testi può naturalmente essere ottenuta attraverso l'utilizzo di un istanza del parser robusto presentato in [3, 5]. Data la rappresentazione in XDG delle frasi componenti il corpus analizzato, la forma normale dei contesti verbale può essere facilmente estrapolata. Infatti, se v è un verbo che appare nella frase S rappresentata da $XDG_S = (C, L)$, la frase contribuisce almeno con il contesto verbale $(v, ((s_1, l_1), ..., (s_n, l_n)))$ dove, prelevando il costituente $c \in C$ che sia retto dal verbo v ovvero $h(c) = v$, per ciascun legame $(c, m) \in L$ di tipo $subj$, $dirobj$ o $dirobj2$ viene posto un (s_i, l_i) con $l_i = concept(m)$ e $s_i = type(c, m)$ e per ciascun legame di tipo preposizionale viene posto un (s_i, l_i) con $l_i = concept(m)$ e $s_i = head(m)$. La funzione $concept(m)$ è il concetto di interesse che sia governatore potenziale (ovvero portatore di significato) per il costituente in questione ovvero:

$$concept(m) = \begin{cases} fc(m) & \text{se } fc(m) \in Conc \\ concept(gov(m)) & \text{se } fc(m) \notin Conc \end{cases} \quad (11)$$

Il modello esteso di dominio può dunque essere facilmente visto attraverso una sua rappresentazione normalizzata che consente l'analisi delle relazioni al fine di poter far emergere quelle che vengono reputate essere le più importanti.

3.2 Contare efficacemente le istanze delle relazioni

La valutazione dell'importanza per il dominio, sia quella basata solo sulla frequenza di apparizione che quella basata sulla misura contrastiva, richiede di contare le istanze di ogni relazione possibile nel corpus. Data la conformazione dei contesti dovuta alla libertà di movimento degli argomenti delle relazioni, il problema non è facilemente risolvibile. Infatti, l'algoritmo semplice di

- prelevare ciascuna istanza del corpus $c \in C(v)$ e riscriverla in una forma linearizzata rispetto a tutte le relazioni $r \in R(v)$
- contare le forme linearizzate attraverso un algoritmo di ordinamento $O(nlog(n)$

è inapplicabile data la cardinalità elevata dello spazio da analizzare. Il numero totale di forme linearizzate da analizzare è nel peggiore dei casi $n = |C(v)||R(v)|$. Se, quindi, si concentra l'attenzione su un solo verbo v si può osservare come il numero delle relazioni possibili $R(v)$ è strettamente dipendente dalla cardinalità dell'insieme $A(v) = A_\Sigma(v) \cup A_\Lambda(v)$ degli argomenti relativi al verbo stesso e dal numero massimo $MC(v)$ degli argomenti che appaiono nei suoi contesti, ovvero:

$$|R(v)| = \sum_{i=1...MC(v)} \left(\begin{array}{c} |A(v)| + i - 1 \\ i \end{array} \right) \tag{12}$$

Dato quindi che $A(v)$ racchiude, tra le altre, le lessicalizzazioni degli argomenti verbali, il numero $|R(v)|$ delle relazioni che devono essere verificate potrebbe essere intrattabile. Il numero totale delle linearizzazioni dei contesti verbali del modello esteso del dominio è decisamente elevato dal momento che deriva da numeri così grandi.

Al fine di poter quindi di poter trattare una così grande mole di dati occorre sviluppare delle euristiche che permettono di arrivare all'obbiettivo di valutare l'importanza per il dominio delle varie relazioni proposte. Le euristiche possono basarsi, da una parte, su alcune caratteristiche della funzione obbiettivo che descrive l'importanza per il dominio e, dall'altra, su caratteristiche specifiche delle informazioni che si trattano.

Osservando in primo luogo che solo le relazioni più importanti sono di interesse, si può evitare di percorrere i rami della ricerca che sicuramente non portano a risultati interessanti. In particolare, supponendo di stabilire come interessanti le relazioni che appaiono più di un valore soglia K, i rami della ricerca che sicuramente portano a valutare relazioni che sono sotto questa soglia di frequenza possono essere abbandonati. Il problema risulta essere dunque quello di individuare le relazioni promettenti per le quali sia possibile sorpassare il limite del valore della soglia di frequenza K.

Dato dunque il vincolo di frequenza sulle relazioni, si possono individuare le strade che non portano a superarlo sfruttando una caratteristica dei contesti verbali in analisi. In particolare, se una relazione $r = (ra_1, ..., ra_m)$ relativa al verbo v ha più di K istanze nel corpus $C(v)$, allora deve sicuramente accadere che la sua versione $\widehat{\Sigma}(r)$ totalmente non lessicalizzata appaia più di K volte nel corpus $C(v)$. La proiezione $\widehat{\Sigma}(r)$ della relazione r sullo spazio sintattico Σ viene definita come:

$$\widehat{\Sigma}(r) = (\widehat{\Sigma}(ra_1), ..., \widehat{\Sigma}(ra_m))$$

dove $\widehat{\Sigma}(ra_i) = ra_i$ se ra_i è un argomento sintattico ($ra_i \in A_\Sigma(v)$) mentre $\widehat{\Sigma}(ra_i) = s_i$ se $ra_i = (s_i, l_i)$ è un argomento sintattico lessicalizzato ($ra_i \in A_\Lambda(v)$). Le relazioni generiche dunque possono essere proiettate sullo spazio sintattico tramite una funzione di associazione. Questa proiezione permette di estrarre e, quindi, di concentrarsi solo su un sottoinsieme $R_\Sigma(v)$ delle relazioni $R(v)$ da cui è possibile estrarre un sottoinsieme $\overline{R_\Sigma}(v)$ promettente che possa guidare la ricerca delle relazioni interessati. Il problema dell'analisi esaustiva dello relazioni possibili nello spazio sintattico è risolvibile perché il numero di delle relazioni da investigare è $|R_\Sigma(v)|$ molto minore di $|R(v)|$ in quanto $|A_\Sigma(v)| = \#preposizioni + 2$ e, quindi, $|A_\Sigma(v)| << |A_\Lambda(v)|$.

Trovato dunque l'insieme $\overline{R_\Sigma}(v)$, si può costruire l'insieme $\overline{R}(v)$ come l'immagine di $\overline{R_\Sigma}(v)$ nello spazio $R(v)$, ovvero:

$$\overline{R}(v) = \{r | \widehat{\Sigma}(r) \in \overline{R_\Sigma}(v)\} \tag{13}$$

Lo spazio dell'analisi si è dunque ridotto poiché l'insieme in questione è di cardinalità inferiore dell'insieme di partenza.

Riscrivendo, quindi, l'algoritmo che ha iniziato il paragrafo, si può ora ridurre il problema ad una conta attraverso un algoritmo di ordinamento O(nlog(n)) su di un numero di esempi più ristretto. Dal momento che lo spazio è stato tagliato in virtù del fatto che lo scopo e quello di trovare le relazioni più frequenti, l'algoritmo di selezione e di conta delle relazioni più importanti può essere quindi formalizzato come segue:

procedure SelectAndRankRelations($R(v)$,$C(v)$)
  begin
    Selezionare $\overline{R_\Sigma}(v) = \{r \in R_\Sigma(v) | hits(r, C(v)) \geq K\}$;
    Sia $L = \emptyset$;
    for each $r \in \overline{R_\Sigma}(v)$
      $L := L \cup prj(C(v), r)$;
      $RankedR(v)$:= CountEquals(L);
    return $RankedR(v)$;
  end

dove $hits(r, C(v))$ è il numero di istanze della relazione r in $C(v)$ e $prj(C(v), r)$ è l'insieme delle contesti proiettati sulla relazione r e $RankedR(v)$ è l'insieme di coppie (f, r) con f la frequenza della relazione $r \in \overline{R}(v)$.

Questo semplice algoritmo permette di calcolare efficacemente, limitatamente alle più frequenti, la frequenza delle relazioni in un insieme di istanze delle stesse. L'algoritmo è generale e risulta essere applicabile a problemi di apprendimento non supervisionato di concetti le cui proprietà siano espresse da feature booleane e in cui sia interessante sapere quale siano le caratteristiche più frequenti in comune tra le istanze dei concetti da scoprire. L'applicazione risulta essere possibile ona volta definita la funzione di proiezione che in questo caso è stata chiamata $\widehat{\Sigma}$.

4 Una ontologia per un sistema di IE: NAMIC, un caso di studio

NAMIC [1, 4] è un sistema per la categorizzazione e per l'*authoring* di un flusso multilingue di notizie provenienti da giornalisti di agenzie di stampa. Per *authoring* si intende la costruzione automatica di hyperlink tra le notizie. Rispetto alle metodologie esistenti di hyperlinking automatico [10], in NAMIC si propone un metodo basato su conoscenza che possa collegare porzioni di notizie molto dissimili e possa giustificare i legami costruiti. Ad esempio, attraverso questo sistema si vorrebbe permettere di collegare notizie seguenti:

1. *Intel, the world's largest chipmaker, bought a unit of Danish cable maker NKT that designs high-speed computer chips used in products that direct traffic across the internet and corporate networks.*
2. *The giant chip naker Intel said it acquired the closely ICP Vortex Computersysteme, a German maker of systems for storing data on computer networks, to enhance its array of storage products.*
3. *Intel ha acquistato Xircom Inc. per 748 milioni di dollari.*
4. *Le dichiarazioni della Microsoft, infatti, sono state precedute da un certo fermento, dovuto all'interesse verso Linux di grandi ditte quali la Corel, Compaq e non ultima Intel (che ha acquistato quote delle Red Hat).*

poichè descrivono tutte attività di acquisizione della *IBM*. Naturalmente, la stessa informazione viene espressa con forme superficiali differenti nelle notizie in lingua omogenea.

L'approccio adottato in NAMIC per il tracciamento dei legami tra le notizie è basato sulla costruzione di una *rappresentazione oggettiva* del loro contenuto. In questa rappresentazione oggettiva (molto simile al *template* dell'IE) vengono espressi, in una forma canonica, i principali eventi descritti nella notizia stessa. Al fine di legare le notizie precedenti per il fatto che riportano tutte acquisizioni delle Intel, devono essere associate nelle rappresentazioni oggettive almeno i seguenti eventi di compravendita (`buy_event`):

1. `(buy_event,(buyer:`**`Intel`**`,bought:`**`a_unit_of_NKT`**`))`
2. `(buy_event,(buyer:`**`Intel`**`,bought:`**`ICP_Vortex_Computersysteme`**`))`
3. `(buy_event,(buyer:`**`Intel`**`,bought:`**`Xircom_Inc.`**`,amount:`**`748000000$`**`))`
4. `(buy_event,(buyer:`**`Intel`**`,bought:`**`quote_della_Red_Hat`**`))`

Le descrizioni degli eventi stessi riconoscibili dal sistema e le loro proprietà osservabili nel testo sono descritte in una ontologia. La rappresentazione di questa ontologia si basa sul formalismo XI [8] adottato in LaSIE [9].

4.1 Analisi di una ontologia acquisita

Lo scopo dell'ontologia sottostante al sistema NAMIC è quello di modellare i domini particolari di applicazione. A differenza dei sistemi "classici" di IE, il sistema non ha a priori un bisogno informativo da soddisfare ma, piuttosto, deve fornire un insieme di legami tra notizie

f	Forma superficiale	DI
2924	last_year	
1739	chief_executive	√
1138	last_week	
1086	next_year	
956	percentNE_stake	√
946	entityNE_share	√
834	last_month	
737	oil_price	
687	joint_venture	√
641	first_half	
631	pre-tax_profit	√
618	interest_rate	√
583	entityNE_yesterday	
575	entityNE_company	√
551	stake_in_entityNE	√
499	prime_minister	√
453	first_time	
438	entityNE_market	√
431	entityNE_index	√
429	earnings_per_share	√
413	share_in_entityNE	√
412	mobile_phone	
396	profit_of_currencyNE	√
374	next_month	
361	second_quarter	
358	entityNE_official	
348	second_half	
341	few_year	
341	same_time	
337	entityNE_government	√
332	next_week	
318	last_night	
316	percentNE_rise	√
316	end_of_the_year	
309	end_of_dateNE	
299	entityNE's_share	√
291	economic_growth	√
285	recent_year	
281	loss_of_currencyNE	√
281	central_bank	√
275	entityNE_deal	√
269	percentNE_increase	√
267	percentNE_stake_in_entityNE	√
248	public_offering	√
240	executive_of_entityNE	√
237	net_profit	√
234	past_year	
234	entityNE_economy	√
230	acquisition_of_entityNE	√
229	entityNE_shareholder	√

Tabella 1: Concetti complessi in $FinTimesNews$

che appartengono ad un determinato dominio conoscitivo. In questa sezione, viene presentata una parte dell'ontologia utilizzata per modellare il dominio finanziario. L'ontologia è stata appresa automaticamente attraverso le tecniche descritte in questo capitolo partendo da una collezione di circa 13.000 notizie finanziarie pubblicate dal Financial Times nel 2000 e 2001. Questa collezione di notizie fornisce il modello esteso di dominio $FinTimesNews$ sul quale applicare l'algoritmo di apprendimento non supervisionato. La funzione di importanza utilizzata è stata la semplice frequenza.

L'analisi del corpus $FinTimesNews$ ha dato origine a 5000 concetti complessi che sorpossano la soglia di importanza (ovvero frequenza di apparizione nel corpus 10). Nella tabella 1 vengono mostrati i 50 concetti complessi che sono ritenuti più importanti nel dominio in questione. Nella tabella viene presentata la forma superficiale del concetto, la sua importanza rappresentata dalla frequenza f e un giudizio dato manualmente sulla sua importanza effettiva per il dominio in questione (*Domain Importance*, DI). Tra i 50 concetti complessi più frequenti, quelli ritenuti effettivamente importanti per il dominio sono 29, ovvero un tasso di precisione del 58%. In particolare si può notare che la maggior parte di quelli non significativi possono essere ricondotti ad una espressione temporale. Questi pur non essendo interessanti per il dominio hanno una grande rilevanza nel riconoscimento di eventi localizzati nel tempo. Tra i 29 concetti interessanti di questa lista, numerosi sono concetti generalizzati poiché

Evento	Descrizione
Acquisition	Compra-vendita di aziende, di porzioni di aziende o di azioni.
Company assests	Risultati finanziari delle aziende in termini di profitti, di dividendi distribuiti, di perdite, variazioni del valore delle azioni, etc..

Tabella 2: Alcuni eventi tipici del dominio finanziario in NAMIC

sottintendono una entità nominata.

La ricerca delle relazioni importanti viene fatta una volta stabiliti quali siano i concetti complessi ammessi nell'ontologia come importanti. In questo caso sono stati scelti come importanti tutti quei 5000 concetti che hanno sorpassato la soglia di importanza.

Anche le relazioni importanti sono state estratte utilizzando una funzione di importanza a frequenza. In particolare, si è decisa come prima soglia per indirizzare l'algoritmo una soglia a frequenza 20. Le relazioni importanti mantenute ed esposte alla validazione manuale sono quelle che, alla fine, superano la soglia di frequenza 15.

L'analisi manuale delle relazioni ritenute interessanti dal sistema ha permesso la definizione degli eventi tipici per il dominio finanziario (in Tab. 2 ne sono riportati alcuni). Queste relazioni coinvolgono i concetti tipici del dominio in questione. Tuttavia, l'associazione delle relazioni tipiche di domini agli eventi è stata fatta in maniera manuale. Ad esempio, le relazioni tipiche associate all'evento *Company assests* sono le seguenti:

(cut,[(subj,entityNE),(dirobj,cost)]))
(rise,[(subj,profit),(to,currencyNE)])
(rise,[(from,currencyNE),(subj,profit),(to,currencyNE)])
(issue,[(subj,entityNE),(dirobj,profit_warning)]))
(suffer,[(subj,entityNE),(dirobj,loss)])
(report,[(subj,entityNE),(dirobj,loss_of_currencyNE)])
(announce,[(subj,entityNE),(dirobj,loss_of_currencyNE)])
(close,[(at,currencyNE),(subj,share)])
(close,[(at,currencyNE),(subj,entityNE_share)])
(rise,[(subj,entityNE_share),(from,currencyNE),(to,currencyNE)])

dove si nota che i concetti coinvolti nelle relazioni sono i profitti, le perdite e le azioni delle compagnie (*entityNE_share*) che salgono e scendono.

Si noti che non tutte le relazioni che possono essere trovate e che sono reputate essere rilevanti attraverso il modello di importanza scelto sono relative al dominio scelto. Si prenda ad esempio la lista delle relazioni governate dal verbo *to make* riportata in Tab. 3. Solo il 30% delle relazioni è considerato interessante in una analisi manuale (colonna *Domain Importance* DI). Tuttavia, la maggior parte delle restanti relazioni, come *make sense for* e *make use* possono essere filtrate attraverso una analisi contrastiva essendo espressioni della lingua generale.

f	**Forma superficiale**	DI
150	*(make,[(dirobj,sense)])*	
132	*(make,[(dirobj,money)])*	✓
121	*(make,[(dirobj,profit)])*	✓
118	*(make,[(dirobj,decision)])*	
108	*(make,[(for,entityNE)])*	
106	*(make,[(dirobj,sense),(subj,null)])*	
102	*(make,[(in,locationNE)])*	
100	*(make,[(to,entityNE)])*	
100	*(make,[(dirobj,null),(for,entityNE)])*	
95	*(make,[(subj,company)])*	✓
87	*(make,[(dirobj,acquisition)])*	✓
83	*(make,[(for,null),(subj,entityNE)])*	
81	*(make,[(dirobj,null),(to,entityNE)])*	
80	*(make,[(dirobj,null),(in,locationNE)])*	
79	*(make,[(dirobj,progress)])*	✓
76	*(make,[(in,entityNE)])*	
75	*(make,[(dirobj,null),(subj,company)])*	✓
71	*(make,[(subj,locationNE)])*	
71	*(make,[(dirobj,use)])*	
71	*(make,[(dirobj,difference)])*	
66	*(make,[(dirobj,use),(of,null)])*	
65	*(make,[(subj,entityNE),(to,null)])*	
60	*(make,[(dirobj,offer)])*	✓
57	*(make,[(subj,null),(to,entityNE)])*	
57	*(make,[(dirobj,null),(in,entityNE)])*	
55	*(make,[(dirobj,profit),(subj,null)])*	✓
55	*(make,[(dirobj,null),(subj,locationNE)])*	
54	*(make,[(dirobj,effort)])*	
53	*(make,[(in,locationNE),(subj,null)])*	
53	*(make,[(dirobj,currencyNE)])*	✓
51	*(make,[(dirobj,mistake)])*	
50	*(make,[(dirobj,null),(subj,entityNE),(to,null)])*	
49	*(make,[(dirobj,debut)])*	✓
48	*(make,[(for,entityNE),(subj,null)])*	
48	*(make,[(dirobj,money),(subj,null)])*	✓
48	*(make,[(dirobj,bid)])*	✓
47	*(make,[(dirobj,locationNE)])*	
46	*(make,[(on,null),(subj,entityNE)])*	
45	*(make,[(dirobj,null),(for,entityNE),(subj,null)])*	
45	*(make,[(dirobj,entityNE),(dirobj2,null),(subj,null)])*	
45	*(make,[(dirobj,difference),(subj,null)])*	
44	*(make,[(dirobj,sense),(subj,it)])*	
42	*(make,[(dirobj,progress),(subj,null)])*	
42	*(make,[(dirobj,decision),(subj,null)])*	
41	*(make,[(dirobj,investment)])*	✓
40	*(make,[(dirobj,payment)])*	✓
39	*(make,[(dirobj,case)])*	
38	*(make,[(dirobj2,currencyNE)])*	
37	*(make,[(dirobj,contribution)])*	
35	*(make,[(with,entityNE)])*	
35	*(make,[(dirobj,loss)])*	✓

Tabella 3: Relazioni rette da *to make* in $FinTimesNews$

5 Conclusioni

In questo articolo è stata presentata l'applicazione delle estrazione terminologica al problema della costruzione del modello del dominio per un sistema di accesso all'informazione testuale. A tale scopo è stato presentato l'adattamento delle tecniche sviluppate per l'estrazione automatica di terminologia. Come si è visto, gli aggiustamenti dipendono principalmente dai domini cognitivi tipici dei sistemi di IE che sono diversi da quelli classici in cui i sistemi di estrazione di terminologia vengono applicati. Si è mostrato come l'applicazione di una tecnica di estrazione terminologica cambia la prospettiva di progetto dei sistemi di IE in quanto il bisogno informativo non è più al centro dell'interesse.

E' stata inoltre mostrata una tecnica originale per l'estrazione di relazioni tra concetti tipiche del dominio. La ricerca delle relazioni tipiche di dominio richiede sia la definizione della forma superficiale di interesse che lo sviluppo di una funzione di importanza. Quest'ultima, non richiesta per le relazioni generiche (ISA, PART-OF), è quella che pone i maggiori problemi computazionali. La metodologia proposta fornisce un algoritmo per il calcolo efficiente della frequenza delle forme superficiali delle relazioni. Il calcolo efficace della frequenza permette la costruzione di funzioni di importanza per il dominio più complesse come quelle basate sull'analisi comparativa tra i domini.

Infine, è stato presentato un caso studio: la costruzione di una ontologia di dominio per il sistema NAMIC.

Si è cercato di mostrare quali possono essere gli effetti benefici dell'applicazione delle metodologie per l'estrazione della terminologia al caso della costruzione automatica delle ontologie soggiacenti ai sistemi di accesso automatico all'informazione testuale.

Riferimenti bibliografici

[1] Roberto Basili, Roberta Catizone, Luis Padro, Maria Teresa Pazienza, German Rigau, Andrea Setzer, Nick Webb, Yorick Wilks, and Fabio Massimo Zanzotto. Multilingual authoring: the namic approach. In *Proceedings of the WORKSHOP ON HUMAN LANGUAGE TECHNOLOGY AND KNOWLEDGE MANAGEMENT, held jointly with ACL'2001 Conference*, 2001.

[2] Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. A contrastive approach to term extraction. In *Proc. of the Conference on Terminology and Artificial Intelligence, TIA2001*, Nancy, France, 2001.

[3] Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. Customizable modular lexicalized parsing. In *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy, 2000.

[4] Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. Web-based information access: Multilingual automatic authoring. In *International Conference on Information Technology: Coding and Computing ITCC 2002 Sponsored by IEEE Computer Society*, Las Vegas, Nevada, 2002.

[5] Roberto Basili and Fabio Massimo Zanzotto. Parsing engineering and empirical robustness. *Natural Language Engineering*, to appear, 2002.

[6] Anne Condamines and Josette Rebeyrolle. Ctkb: A corpub-based approach to terminological knowledge base.

[7] L. Davidson, J. Kavanagh, K. Mackintosh, I. Meyer, and D. Skuce. Semi-automatic extraction of knowledge-rich contexts from corpora.

[8] Robert Gaizauskas and Kevin Humphreys. Xi: A simple prolog-based language for cross-classification and inheritance. In *Proceedings of the 7th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA96)*, Sozopol, Bulgaria, 1996.

[9] Robert Gaizauskas and Kevin Humphreys. Using a semantic network for information extraction. *Natural Language Engineering*, 3, Parts 2 & 3:147–169, 1997.

[10] Stephen J. Green. *Automatically generating hypertext by computing semantic similarity*. PhD thesis, Department of Computer Science, University of Toronto, 1997.

[11] Emmanuel Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. PhD thesis, Univesité de Nantes, Faculté des Sciences et de Techniques, 1999.

[12] MUC-6. Proceedings of the sixth message understanding conference(muc-6). In *Columbia, MD*. Morgan Kaufmann, 1995.

[13] MUC-7. Proceedings of the seventh message understanding conference(muc-7). In *Columbia, MD*. Morgan Kaufmann, 1997.

[14] Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, Portland, Oregon, 1996.

[15] Ellen Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, 1999.

[16] Roman Yangarber. *Scenario Customization for Information Extraction*. PhD thesis, Courant Institute of Mathematical Sciences, New York University, 2001.

[17] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Unsupervised discovery of scenario-level patterns for information extraction. In *Proceedings of Conference on Applied Natural Language Processing ANLP-NAACL 2000*, Seattle, WA, 2000.