Università degli Studi di Salerno

Dipartimento di Scienze Politiche, Sociali
e della Comunicazione

Applicazioni di NLP per il Semantic Web

Trattamento delle polirematiche terminologiche nel dominio della Medicina

A. Elia, F. Marano, M. Monteleone, J. Monti, A. Napoli, A. Postiglione, D. Vellutino

Outlines

- Obiettivi della Ricerca
- Problema del Semantic Web: ontologie e NLP come possibili soluzioni
- Rapporto tra terminologia e poliremathiche
- Metodologia e strumenti
- Case study: applicazioni di NLP nel dominio terminologico della Medicina
- Risultati: prodotti terminologici (dizionari elettronici, grammatiche locali, corpora)
- Conclusioni e lavori futuri

Obiettivi della Ricerca

- Migliorare le performances dei sistemi informativi, di Knowledge Management (motori di ricerca) per promuovere lo sviluppo del Semantic Web
  - Affrontando i problemi linguistici
  - Mettendo in luce l'importanza del rapporto tra terminologia e trattamento automatico del linguaggio
  - Proponendo dei modelli teorici e metodologici per la costruzione di applicazioni di Natural Language Processing
  - Sviluppando applicazioni NLP

Semantic Web… Utopia?

- La più grande ed utopica visione del SW prevede:
  - La grande mole di dati disponibili sul web dovrà essere organizzata in modo tale da permettere di indicizzare i documenti in base al loro **significato** e non in base alla "forma" dei contenuti (Allemang, 2006).
    - Il primo passo è analizzare il linguaggio dato che è lo strumento con cui un utente interroga i sistemi di gestione delle informazioni.

Ontologie: una possibile soluzione per SW

- Dal punto di vista del SW le ontologie possono essere definite come:
  - un documento che esplicita una relazione tra termini (Berners-Lee, 1991);
  - la semplificazione di una concettualizzazione e quindi un'astratta e semplificata visione del mondo che vogliamo descrivere (Gruber, 1993).

Forma di classificazione

NLP e i problemi linguistici del SW

- Gli attuali e più diffusi studi sul SW tendono a privilegiare il problema sull'adozione di standard condivisi tralasciando invece l'ostacolo più grande.
  - Problema linguistico
- Studi di NLP sono molto rilevanti per andare incontro alle esigenze del SW dato che gli ostacoli linguistici interferiscono molto col SW.
  - Sviluppo di ontologie su base linguistica

Problema linguistico

1. Formulare la query usando frasi in linguaggio naturale.
   - Generalmente l'utente deve sforzarsi di tradurre le query in keyword.
2. Filtrare documenti scegliendo solo quelli pertinenti.
   - Lo sviluppo di ontologie su basi lessicali, sintattiche e semantiche porterebbe ad un miglioramento dei risultati di Information Retrieval.

- Considerato questo contesto, un'analisi approfondita e una formalizzazione esaustiva del lessico permettono:
  - Gestione delle informazioni
  - Attività di categorizzazione e classificazione attraverso esplicitazione di etichette semantiche

Tipi di combinazioni lessicali: libere e ristrette

- LG ha individuato le condizioni di continuum per le combinazioni lessicali in sintagmi o frasi, che possono essere di quattro tipi:
  - distribuzione **libera** = **elevata** variabilità di co-occorrenza fra le parole con significato composizionale e denotato.
    - Es. verbo > sentire (dolore, musica, Max)
    - Es. nome > garza (bianca, pulita, insanguinata)
    - Es. avverbio > con (dolore, fatica)
  - distribuzione **ristretta** = **ridotta** variabilità di co-occorrenza fra le parole
    - Es. verbo > medicare (ferita, gamba, Max)
    - Es. nome > garza (sterile, elastica, adesiva)
    - Es. avverbio > da un (momento, giorno, anno..) all'altro

Tipi di combinazioni lessicali: fisse e invariabili

- distribuzione **fissa** = **nulla** o quasi nulla variabilità di co-occorrenza fra le parole.
  - Es. verbo > dire trentatrè
  - Es. nome > malattia esantematica, agente patogeno, cellula staminale
  - Es. avverbio > chiaro e tondo
- Sono combinazioni lessicali a distribuzione fissa le **polirematiche** e le **frasi idiomatiche**. Entrambe sono sintagmi con atomicità semantica (**unità di significato**).
- struttura **invariabile** = senza alcuna variabilità di co-occorrenza fra le parole, come nel sintagma dei proverbi
  - Es. Chi rompe paga e i cocci sono i suoi

Importanza delle poliremetiche nella terminologia

- La descrizione del lessico è più precisa se governata dal trattamento delle poliremetiche.
  - Dall'osservazione sui testi emerge una massiccia presenza di poliremetiche.
- Nell'ambito di LG lo studio del lessico terminologico riguarda l'uso delle unità lessicali superiori classificabili come poliremetiche, vale a dire un "gruppo di parole che ha un significato unitario, non desumibile da quello delle parole che lo compongono, sia nell'uso corrente sia nei linguaggi tecnico-specialistici", come indicato dal dizionario di De Mauro (2000).
- Le poliremetiche sono strettamente legate alla terminologia perchè essa, da un punto di vista formale e semantico sfrutta le procedure di formazione delle parole composte.
  - Nome come "sistema" dal significato generico può essere maggiormente specificato aggiungendo altri elementi lessicali:
    - Sistema nervoso
    - Sistema nervoso centrale
    - Sistema nervoso periferico

Metodologia: Lessico-Grammatica

- LG è una teoria basata su modelli matematici del linguaggio (Bloomfield, 1933; Harris, 1982; Schützenberger in Gross et al., 1973).
- Le teorie LG si sono sviluppate a partire dagli anni '60 grazie al linguista francese Maurice Gross (Gross, 1968; 1989).
- Diversamente da teorie formaliste meglio conosciute e basate sulla sintassi (Chomsky, 1957; 1965), LG ritiene che la descrizione del linguaggio debba partire dall'osservazione del lessico e del comportamento distribuzionale delle entrate lessicali, fondendo così la sintassi e il lessico.
  - http://infolingu.univ-mlv.fr/ (click on "Bibliographie")
  - http://en.wikipedia.org/wiki/Operator_Grammar
  - http://en.wikipedia.org/wiki/Zellig_Harris
  - http://fr.wikipedia.org/wiki/Lexique-grammaire
  - http://it.wikipedia.org/wiki/Lessico-grammatica

Strumenti

- Principali strumenti sviluppati sono:
  - Dizionari elettronici
    - Parole semplici: 135 mila (1.200 mila forme flesse)
    - Parole composte (dizionari elettronici terminologici): 154 mila (480 mila forme flesse)
  - Tavole/matrici lessico-grammaticali
    - Verbi, nomi predicativi, verbi supporto
    - Frasi idiomatiche
  - Grammatiche locali
    - Automi a stati finiti
    - Trasduttori a stati finiti

Case study: il corpus

- È stato costruito un monitor corpus nel dominio della Medicina collezionando testi da un manuale di medicina edito da Merck Sharp & Dohme, disponibile on line http://www.msd-italia.it/altre/manuale/index.html.

	Token	Type
Corpus MED	899048	36370

- Su 5858 polirematiche riconosciute da NOOJ, software di NLP, ben 3913 (66%) sono specifiche di dominio.

Case study: corpus annotato in XML

- NOOJ ha automaticamente individuato ogni polirematica e l'ha annotata in XML con tag semantici del dominio della Medicina.
  - I <LU LEMMA="meccanismo di difesa" CAT="N" FLX="C7" Genere="m" Numero="p" Class="NPN" Term="MED">meccanismi di difesa</LU> includono le barriere naturali (p. es., la cute e le mucose) le <LU LEMMA="risposta immune" CAT="N" FLX="C544" Genere="f" Numero="p" Class="NA" Term="MED">risposte immuni</LU> aspecifiche (p. es., cellule fagocitarie [neutrofili, macrofagi e i loro prodotti); e le <LU LEMMA="risposta immune" CAT="N" FLX="C544" Genere="f" Numero="p" Class="NA" Term="MED">risposte immuni</LU> specifiche (p. es., anticorpi).

Case study: text classification

- Il corpus è stato processato da un altro software, CATALOGA, classificatore di testi terminologici in base al campo semantico.

Dominio di conoscenza	Polirematiche terminologiche
Medicina	76.47 %
Economia	4.99 %
Informatica	3.02 %
Diritto	2.51 %
Fisica	1.09 %

Dizionari elettronici terminologici

- Dizionari elettronici terminologici per 180 campi semantici. I più importanti sono:
  - Informatica (54.000 entrate)
  - Medicina (46.000 entrate)
  - Diritto (21.000 entrate)
  - Ingegneria (1.9000 entrate)
  - …

Stringa del dizionario elettronico bilingue della Medicina

```
ubriachezze patologiche, ubriachezza patologica, N + Genere =
f + Numero = p + Class = NA + Term = MED + Eng = pathologic
intoxications, pathologic intoxication, Number = s+ Class = AN
```

Polirematiche terminologiche

- Riconoscimento, formalizzazione, ontologizzazione

Struttura interna	Entrate	% sul totale (5,858)	Entrate MED	% sul tot MED (3,913)
NA	4,089	69.80	2962	75.70
NPN	1,425	24.33	818	20.90
NN	157	2.68	108	2.76
AN	153	2.61	25	0.64
Others (Avv., Prep., etc.)	34	0.58	/	/

Risultati: prodotti terminologici

- Dizionario poliremathiche medicina: 46.000 entrate lessicali
- Monitor Corpus medicina: 899.048 parole
- Monitor Corpus annotato con tag XML

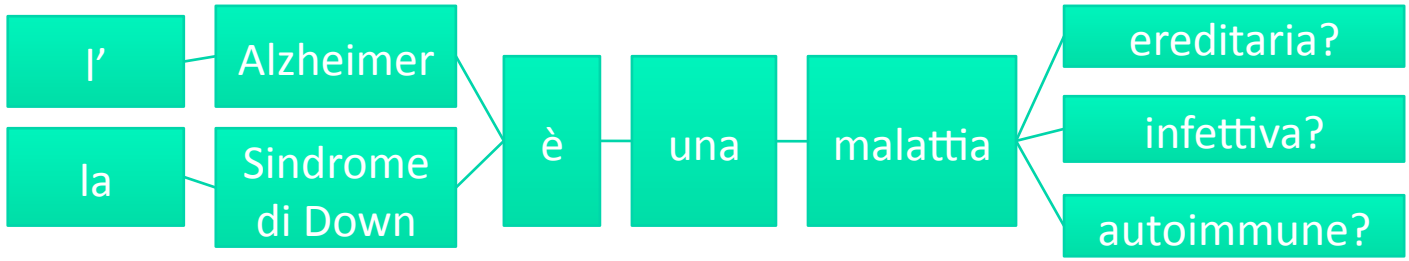
Sviluppo di Ontologie di dominio su base linguistica

Conclusioni

- I risultati confermano la stretta relazione che c'è tra polirematiche e terminologia.
- Una coerente formalizzazione del linguaggio produce risorse linguistiche.
- Risorse linguistiche sono di Alta Qualità perché estratte manualmente grazie alla competenza dei linguisti e validate da esperti di dominio.
- Risorse linguistiche possono essere utilizzate per sviluppare "efficienti ed efficaci" applicazioni di NLP:
  - Information Retrieval, Question Answering
  - Information Extraction
  - Ontologie
  - Machine Translation

Lavori futuri

- Aggiornamento dei dizionari terminologici testing su corpora
- Aggiornamento del corpus
- Sviluppo di grammatiche locali in forma di automi/ trasduttori per descrivere specifici fenomeni del linguaggio
  - Realizzare risponditore automatico di query



Grazie

Annibale Elia	elia@unisa.it
Federica Marano	fmarano@unisa.it
Mario Monteleone	mmonteleone@unisa.it
Johanna Monti	jmont@tin.it
Antonella Napoli	antnapoli@unisa.it
Alberto Postiglione	apostiglione@unisa.it
Daniela Vellutino	dvellutino@unisa.it