Descrizione dei criteri e dei metodi per la costruzione di un database di terminologia.

Rita Marinelli

Abstract

Vengono qui descritte le fasi della costruzione di un database di terminologia appartenente al dominio del *turismo,* considerando in particolare un insieme di termini che riguardano la ristorazione e le strutture ricettive. Questo rapporto si articola in tre parti: la prima contiene una descrizione dello studio di fattibilità e il progetto del database di tipo semantico lessicale, in cui ogni termine è visto come punto nodale di una rete linguistico-concettuale; la seconda contiene, sotto forma di slides in power-point, una rappresentazione schematica di esempi di termini e di relazioni per la loro codifica nel database; la terza contiene la relazione sull'attività svolta dall'ILC per la creazione del database.

1. La costruzione del database di terminologia riguardante il *turismo*

1.1 Inserimento dei Concetti

I concetti da inserire come entrate del database sono quelli del campo di conoscenza che è anche campo di lavoro. Ogni concetto è inserito in un campo semantico, con una rete di relazioni che lo collegano ad altri concetti e che vanno rappresentati nel database.
E' per mezzo di relazioni semantiche che il significato di una parola è visto come un "nodo di accesso" nella rete di conoscenza (Langacker, 1987).
Un altro presupposto importante è che i termini devono essere trattati "come entità linguistiche analoghe alle altre unità lessicali rispetto alla loro natura referenziale e alla loro funzione nel discorso" (Cabré, 1998/99), e che si presuppone il riferimento alla semantica cognitiva nella nostra prospettiva nel trattamento dei termini, del loro significato, della loro codifica nel database terminologico, nella struttura concettuale stessa del database.
Il primo passo è quello di fissare i concetti principali da cui partire.
Si possono considerare tre alternative, che non si escludono l'una con l'altra:
a) fissare i concetti principali dell'ontologia e poi popolare il db;
b) popolare il db con concetti (entrate) anche con alto grado di specificità, e poi risalire le catene tassonomiche alla ricerca di concetti TOP;
c) usare una strategia "mista".
La terza, c), da un punto di vista operativo, è quella più adatta alla creazione di un nuovo database di dominio. Infatti viene composta, insieme con i committenti, esperti di dominio e utenti, una lista di concetti interessanti, centrali per il dominio e altamente rappresentativi, provenienti da fonti diverse. Vanno considerate come fonti:

a) i dati forniti dai committenti
b) glossari

c) IWN, ma solo nella fase iniziale
d) altre risorse.

Si può usare la rete generica IWN per vedere se e quali concetti di questa lista sono presenti in IWN; quelli che sono presenti possono essere esportati in un file xml e inseriti nel db specialistico importando il file xml.
Le modalità di import/export sono soggette a vincoli, che qui non specifichiamo, ma che vanno tenuti presenti ("rumore" importando relazioni che connettono termini troppo generici e/o che non c'entrano, ecc.).
Altri concetti devono essere inseriti, appartenenti al dominio D, non presenti in IWN, perché con un alto grado di specificità, ritenuti anch'essi rappresentativi del D per motivi vari, per es. perché hanno un grande numero di iponimi.

2. Le relazioni per rappresentare i concetti

Si devono trasformare in relazioni semantiche le connessioni logiche e soprattutto pragmatiche che ci sono fra i concetti che interessano, e si devono definire chiaramente questi rapporti.
Si individuano gli eventi, le azioni, le entità concrete, ecc. con cui ha a che fare un turista, soprattutto pensando ai "servizi" di cui può avere bisogno. Ciascuno di essi è rappresentato da un concetto che corrisponde a un'entrata del db. I vari componenti del campo semantico di ogni entrata sono nodi di informazione che va sviluppata e popolata.

I contenuti informativi devono essere organizzati con criteri WordNet-like, cioè utilizzando, almeno nella fase iniziale, la struttura concettuale di IWN (basata su relazioni semantico-lessicali).
Si fissano dei concetti fondamentali, per esempio: Commercio, Turismo, Servizi.
Si importano i concetti dal db generico a quello nuovo per cominciare a popolarlo.
Vanno fissati i Broader Terms (BT) o Base Concepts (BC) della terminologia e il modo in cui si sviluppano le catene di significati:

1. in senso gerarchico tassonomico, usando relazioni verticali (ipo/iperonimia);
2. in senso orizzontale (relazioni di parte, di luogo, causa, ecc.) sfruttando relazioni esistenti ove possibile;
3. aggiungendone altre ad hoc:
    a. per la definizione e classificazione di eventi in senso spazio-temporale (orari, per es.), che sono totalmente mancanti;
    b. per l'uso di NP che è prevedibile che saranno molto usati;
4. usando relazioni associative (Related Term o RT, per usare la terminologia ISO (ISO 2788-1986)), fra termini che non sono né sinonimi, né in relazione gerarchica e che rappresentino il legame all'interno del db in modo da suggerire a chi accede al 1° termine anche il 2° (es.: ristorante - menù). RT che è qualcosa di simile alla fuzzynym, già presente in IWN, ma meno generica (e con reciprocità).

Un concetto del Dominio sarà così definito da un insieme di relazioni all'interno del database terminologico; queste relazioni diranno sia qual è il suo iperonimo, sia quali sono gli altri concetti con cui è collegato.
Considerando le catene tassonomiche, i concetti che hanno una posizione "intermedia" sono quelli più rappresentativi del Dominio (Rosch 1978-88).

Consideriamo i concetti nel frame C (Commercio), T (Turismo), S (Servizi).
Le relazioni devono essere del tipo "1 a 1", "1 a molti", "molti a molti".

Si costruiscono le strutture per lo più sulla base di instances e di classi, cioè su NP, ma non solo.
Le relazioni già usate di solito (iponimia, appartenenza, luogo, ecc.) possono essere utilizzate, ma devono essere riorganizzate in modo tale da rimuovere o correggere certi "constraints" o vincoli che attualmente ci sono e che, in questo caso, sarebbero di ostacolo se restassero così come sono. Ad esempio, la relazione "co_agent_result" vale tra "pittore" e "dipinto", perché sono entità del 1° ordine, vale a dire concrete, ma non è applicabile fra "Organizzazione Marittima Internazionale (IMO)" e "codice IMO", perché sono due entità del 2° ordine.

3. Ereditarietà dell'ontologia

La costruzione dell'ontologia è fondamentale per scopi pratici
Anche le specifiche di D devono essere fornite dai committenti: su esse vanno individuate le primitive C, T, S (dal punto di vista ontologico) e un insieme di core concepts ad esse legati per mezzo di relazioni associative (RT) da concordare. I core concepts sono i concetti da cui possiamo e dobbiamo sussumere tutti gli altri.
La relazione di sussunzione che è codificata nella ontologia specifica in termini di iperonimia, deve essere ereditata dagli iponimi: se *trattoria* è sotto *Ristorazione (R)*, *trattoria sul mare* deve essere sotto *Ristorazione*.
Bisogna decidere se la catena ipo/iperonimica è l'unica tramite la quale si trasmette l'ereditarietà dell'ontologia, o se questa ereditarietà si può trasmettere, in certi casi, anche attraverso la relazione di parte. Ora come ora sembrerebbe di sì: se *menù* è sotto R, e *secondi piatti* è parte di *menù*, anche *secondi piatti* è sotto R.

4. Strumenti necessari

I concetti del db sono collegati da relazioni non più 'solo' semantico lessicali, ma funzionali, con caratteristiche di immediatezza e brevità, legate allo scopo pratico. Solo una parte di esse sarà ereditata da IWN; un insieme nuovo andrà pensato "ad hoc", considerando, per es. i vari punti di snodo che in IWN non ci sono, la codifica spazio-temporale, e RT di vario genere da definire in dettaglio con i committenti, esperti di dominio, ma soprattutto con gli **utenti.**

4.1 DB terminologico

C'è bisogno di un db terminologico vuoto, pronto, per fare prove.
Utilizzando il tool, si deve poter aggiornare il set di relazioni semantiche attualmente disponibile e il file dei constraints.
Sarebbe da valutare se cambiare il nome che hanno ora le relazioni.
L'ontologia specifica deve poter essere ereditata.
Inoltre si può pensare ad una lista di nuove entrate da mettere in un file, di tipo xml o txt, per poi poter importare automaticamente.
Il database è fortemente basato su Named Entities, perciò le relazioni che attualmente codificano i NP in IWN devono essere affinate, studiando la possibilità di ereditare l'ontologia (TO) di IWN da quella della classe di appartenenza.
Tra le funzionalità dell'interfaccia informatica per l'interrogazione del database lessicale, deve essere prevista la ricerca per sottostringhe, es.: "Rist*", per ricevere in risposta: "ristorante, ristorazione, ristoranti", ecc.

4.2 Giovani

Particolare attenzione dovrebbe essere data alle risposte che si possono dare alle richieste di informazioni da parte di un "pubblico giovane", alla ricerca di PUB, Paninoteche, Discoteche, Feste in genere, Eventi vari, Punti di Ritrovo, Aperitivi. Si dovrebbe poter rispondere alle domande: DOVE, COME, QUANDO.
Sarebbe interessante raccogliere un insieme di parole che appartengono al gergo dei più giovani, usate per indicare i locali frequentati (disco, pubbino, ecc..) e/o per definirli (figo, casino, ecc.), con riferimento all'uso di forme abbreviate per gli SMS che i giovani si mandano.
Per esempio, considerando un **Centro Commerciale** dal punto di vista dei **servizi** che offre, un primo elenco di concetti con un primo livello di strutturazione potrebbe essere questo:

Autonoleggio
- **Avis**
- **Hertz**

Parcheggio
- Coperto
- Sorvegliato
- ….

Bancomat
SNAI
Supermercato
Edicola Giornali
Informazioni
Punto Internet
Biglietteria
- Linee
- Partenze
- …..

Nursery
Tabacchi
Telefonia fissa e mobile
Ristorazione
- Self Service
- Bar
- Pizzeria
- ……

Negozi
- Ottica
- Sport
- Mare
  - Canotti
  - Pinne
  - Costumi
  - ….
- Abbigliamento
  - NP negozio1
  - NP negozio2
- Prodotti tipici
- Artigianato
- …….

Sono da considerare anche gli insiemi di dati da inserire come “punti di snodo”, cioè dati che possono far parte di più di un campo semantico, vale a dire: gli indirizzi, i numeri di telefono, i giorni di chiusura, e altri insiemi di dati che fanno parte di un determinato campo semantico, per esempio i tipi di menù (vegetariano, carne, pesce, ecc.).
Qui si seguito sono disponibili le slides in power-point con una rappresentazione schematica di esempi di termini e di relazioni da usare per la loro codifica nel database.

Riferimenti

Cabré M.T. 1998/1999. *Do we need an autonomous theory of terms?*, in: «Terminology», vol. 5, n. 1.5-19.

Cabré Castelvì M. T. 2000. *La terminologia: representacion y comunicacion,* Barcelona: IULA.

Langacker R. W., Foundations of cognitive grammar, vol. I: theoretical prerequisites. Stanford, Calif.: Stanford University Press, 1987.

Marinelli R., Spadoni G. 2007. Modelling a Maritime Domain Ontology. *Proceedings of the X International Symposium on Social Communication*. Centre for Applied Linguistics. Santiago de Cuba: VOL. 1. 511-515.

Marinelli R. 2008. Enhancing a Terminological Database with Terms from a Scientific Domain. *Proceedings of the Third BalticConference on Human Language Technologies, October 4-5 2007*, Kaunas,Lithuania, Kaunas, Vytauto Didþiojo Universitetas.165-172.

Rosch, E. Principles of Categorization (1978). In Readings in Cognitive Science, a Perspective from Psychology and Artificial Intelligence, A. Collins & E. E. Smith, Morgan Kaufmann Publishers, San Mateo, California, 1988.

Termini e tipi di relazioni

esempi

Tipi di relazioni

uno a uno esempio: indirizzo, telefono

un numero di telefono è collegato con un solo utente

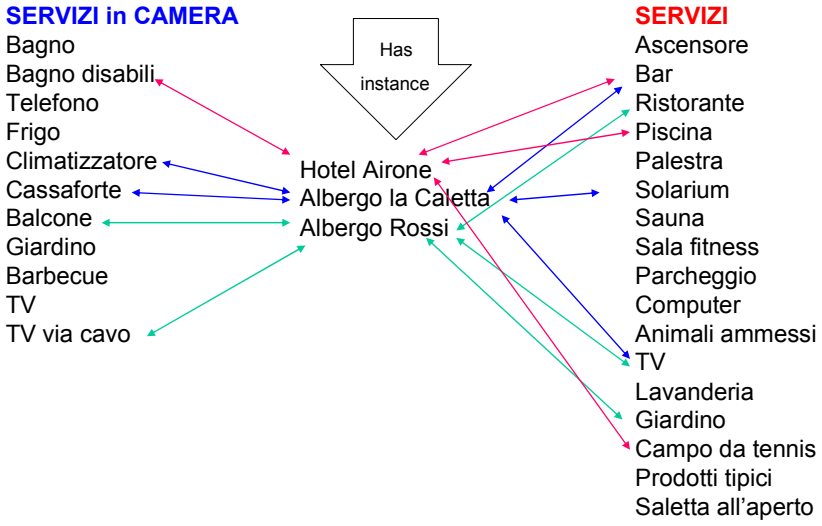
molti a molti esempio: menù, servizi

tanti tipi di menù sono collegati a tanti ristoranti

uno a molti esempio: città, zona

In una stessa zona della città ci possono essere più ristoranti, alberghi, ecc.

Hotel



Hotel

Hotel Airone → *offre* → Piscina, Campo da tennis, Bar

Albergo La Caletta → *offre* → Bar, Solarium, TV

Albergo Rossi → *offre* → Bar, Giardino, TV

Hotel

Viceversa:

Bar → **è offerto da** → Albergo La Caletta, Albergo Rossi, Hotel Airone

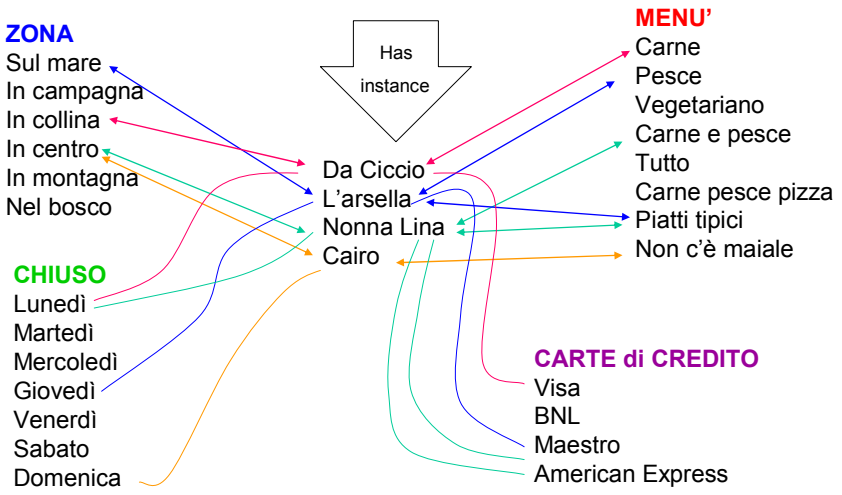
Solarium → **è offerto da** → Albergo La Caletta

Giardino → **è offerto da** → Albergo Rossi

TV → **è offerto da** → Albergo Rossi, Albergo La Caletta

Piscina → **è offerto da** → Hotel Airone

Ristorante



Ristorante

In centro → trovi → **Nonna Lina**, **Cairo**

Nonna Lina → si trova → in centro

Lunedì → chiude → **Da Ciccio**, **Nonna Lina**

Cairo → è chiuso → domenica

Piatti tipici → si mangia da → **L'Arsella**, **Nonna Lina**

L'arsella → puoi mangiare → **Pesce**, **Piatti tipici**

L'arsella → accetta → Maestro

Maestro → è accettata da → **Nonna Lina**, **L'Arsella**

Attività svolta dall'Istituto di Linguistica Computazionale (ILC) del C.N.R. di Pisa.

Da parte dei committenti (ASCOM Grosseto) sono stati ricevuti dei documenti in cui, dopo un'introduzione in cui si descrivono brevemente località, panorami, prodotti della Maremma, viene anche fornito un elenco di luoghi di notevole importanza per il tema del Turismo, per esempio: musei, siti archeologici, ecc.

Viene anche fornito un elenco di strutture ricettive, di stabilimenti balneari, di agenzie di viaggi e di centri commerciali del territorio.

A partire da questi dati, si sono ricavate le tipologie più frequentemente usate di strutture ricettive e le caratteristiche comuni che fossero sufficientemente rappresentative. Analogo procedimento è stato seguito per la parte della ristorazione.

Ogni concetto che rappresenta una caratteristica "prototipica" sia delle strutture ricettive che della ristorazione è stato preso in considerazione e studiato con lo scopo di essere rappresentato poi, nel database di dominio turistico, da un campo preciso di informazione.

Tale campo è destinato a essere non isolato, rigidamente accompagnato da altri campi di una scheda anagrafica, ma anzi è visto come un nodo di accesso a una rete semantica in cui è inserito.

Esso è quindi collegato ad altri campi, che rappresentano, cioè, altre informazioni, e questo avviene per mezzo di relazioni semantico - lessicali.

Le relazioni semantico – lessicali sono state prese in considerazione in riunioni dedicate alla discussione di questo argomento. E' stata fatta un'analisi delle relazioni dei database WordNet – like, che hanno, cioè, per modello (a livello concettuale e pragmatico) sia WordNet, la rete americana di Princeton, sia EuroWordNet, la rete che è il risultato dell'omonimo progetto europeo, sia ItalWordNet, la rete semantico – lessicale sviluppata dall'Istituto di Linguistica Computazionale del C.N.R. di Pisa.

E' stato anche preso in esame il database di terminologia marittima Mariterm (settore tecnico – nautico e dei trasporti marittimi) costruito su modello IWN, per il recupero di risultati dell'esperienza già avuta nel campo della creazione di un database semantico di terminologia di dominio.

Le relazioni del modello di partenza EWN/IWN sono state elencate, analizzate e valutate, ai fini dell'utilizzo delle stesse per la creazione del nuovo database di Turismo.

Sono stati considerati e studiati anche altri modelli di relazioni, questa volta appartenenti al mondo della documentazione che fa uso di standard internazionali per la rappresentazione, la gestione e il recupero dell'informazione (modelli ISO).
E' stata focalizzata così l'attenzione sulla relazione associativa che è in grado di rappresentare connessioni molto strette di similarità di significato fra termini che appartengono a categorie logiche diverse (per esempio: Ristorante – menù).
Normalmente la relazione associativa lega termini che appartengono a categorie logiche diverse. In questo caso svolge anche una importante funzione di collegamento tra strutture gerarchiche diverse. (nelle norme ISO 2788, però, si citano anche casi in cui i due termini appartengono alla stessa categoria o più precisamente sono termini affini ma con significati sovrapposti, per esempio: "navi" e "barche"). Spesso, però, proprio perché non esiste una regola rigida per la sua applicazione, si corre il rischio di utilizzarla in modo inappropriato e soggettivo.
Rispetto alla relazione gerarchica (che in un database semantico come IW è rappresentata dall'iperonimia/iponimia e dalla relazione parte/tutto), la relazione associativa non rappresenta il significato del termine da un punto di vista logico-strutturale ma permette di recuperare la semantica "locale" che, altrimenti, andrebbe perduta durante la classificazione del termine fatta secondo modalità "classiche".

Si sono studiate anche altre relazioni che nel modello IWN non sono presenti, da usarsi per poter fissare e catalogare la situazione spazio – temporale di ciò che avviene e che può interessare al turista: eventi culturali, rappresentazioni teatrali, ecc., ma anche appuntamenti, ritrovi, aperitivi, concerti di musica di tutti i tipi, ecc.
Un'attenzione particolare è stata data alle relazioni che nel modello IWN servono per codificare i Nomi Propri: *has_instance* e *belongs_ to_ class*; i Nomi Propri e le relazioni che li codificano costituiscono la parte predominante fra gli "items" del database di dominio turistico.
Riunioni si sono svolte per studiare in particolare la connessione dei NP con l'ontologia fondazionale di IWN e con quella specifica di dominio, in modo da permettere l'ereditarietà della Top Ontology di IWN non solo tramite la catena tassonomica delle relazioni di iperonimia/iponimia, ma anche tramite le relazioni di appartenenza.
Inoltre per ogni nodo semantico rappresentato da ogni termine di dominio, è stato previsto un collegamento *plug-in* alla rete generica IWN. Esso permette di visualizzare informazioni relative al termine così come sono contenute nel database terminologico, e di vedere le informazioni relative ai nodi alti delle tassonomie che sono nel database generico. Insomma,

grazie a questa relazione di “innesto”, si garantisce coerenza e consistenza del contenuto semantico di un termine, sia dal punto di vista del lessico generale, dia dal punto di vista del lessico specializzato.

Alla luce del materiale esemplificativo fornito, si sono enucleate le tipologie di domande (queries) rappresentative di situazioni tipiche: la ricerca di numeri utili, di informazioni relative a eventi, iniziative culturali, ecc., di strutture ricettive e di ristorazione, di strutture commerciali, di situazioni meteorologiche, pensando sempre in funzione di possibili bisogni da parte di gruppi di turisti.
Si è convenuto su un’architettura concettuale che permetta il crosschecking dei dati; cioè che permetta, per esempio per quanto riguarda la ristorazione, la ricerca generica sui ristoranti, ma che dia anche la possibilità di scegliere un ristorante sulla base della zona in cui è situato, del tipo di menù previsto, del giorno di chiusura, della carta di credito accettata, ecc.
Un primo nucleo di struttura concettuale è stato progettato, come base per futuri ampliamenti, instanziazioni di relazioni, sviluppi e incrementi di termini codificati.
