Ontology learning e semantic annotation (ILC)

Fin dalle fasi iniziali del processo di creazione del libro liquido “nativo digitale”, la piattaforma editoriale iSLe mette a disposizione dell'autore la possibilità di accedere su base semantica ai contenuti presenti nelle collezioni digitali archiviate; ciò significa che i contenuti sono accessibili attraverso modalità di navigazione “intelligente”, grazie alla loro precedente organizzazione in strutture logico-concettuali specializzate per il dominio dei materiali didattici. Tale strutturazione è funzionale all'elaborazione di una sorta di “mappa concettuale” che viene estratta automaticamente dalle risorse digitali esistenti e può essere specializzata per le diverse discipline, così da poter supportare il redattore nella scrittura di un nuovo testo in formato “liquido” per la disciplina di interesse.

Il processo di creazione e gestione semantica dei contenuti in iSLe ha contemplato l'integrazione di alcune delle funzionalità avanzate di estrazione automatica di conoscenza dal testo, che sono state sviluppate dall'Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC) del CNR di Pisa. In particolare, il punto di partenza è costituito dallo strumento T2K² (*Text-to-Knowledge*), una piattaforma software ibrida realizzata congiuntamente dall'ILC-CNR e dall'Università di Pisa, che integra tecnologie avanzate di *Natural Language Processing* (NLP), tecniche statistiche e algoritmi di *machine learning*, con l'obiettivo di trasformare le conoscenze implicitamente codificate all'interno di corpora di dominio, in una rete di conoscenza esplicitamente strutturata. T2K² consente tanto la creazione semi-automatica di repertori terminologici e la loro organizzazione in strutture ontologico-concettuali, quanto l'annotazione semantica dei documenti sulla base dell’ontologia di riferimento.

A partire dal risultato dell'annotazione linguistica del testo a livelli di complessità crescente, lo strumento è in grado di identificare i “correlati” linguistici dei concetti rilevanti presenti nel testo, siano essi identificatori di entità di dominio (espressi come unità lessicali mono- o polirematiche), siano essi identificatori di entità nominate (ovvero nomi propri riferiti a persone, luoghi, organizzazioni e unità geopolitiche). Il passo successivo all'estrazione terminologica è dato dalla strutturazione delle entità riconosciute come rilevanti in un'ontologia di dominio, intesa come una descrizione formale complessa dei concetti di un dominio e delle relazioni che sussistono tra di essi. A questo proposito, un primo livello di strutturazione ontologico-concettuale del dominio, disponibile in T2K², è dato dall'organizzazione dei termini rilevanti in frammenti di catene tassonomiche, ovvero relazioni gerarchiche di iperonimia-iponimia basate su una relazione di inclusione lessicale. Ad esempio, se ipotizziamo che il sistema abbia riconosciuto come rilevanti, all'interno di un testo di storia dell'arte, le unità polirematiche “arte gotica” e “arte classica”, queste ultime saranno classificate come iponimi della testa lessicale nominale “arte”. Ad un livello di strutturazione semantica più avanzato, le entità di dominio e le entità nominate sono messe in relazione tra loro, sulla base del contesto di co-occorrenza o di relazioni di similarità semantica.

Ai fini di personalizzare l’accesso su base semantica alle collezioni documentali considerate nel progetto, le metodologie di *Ontology Learning* e *Semantic Annotation* disponibili in T2K² sono state potenziate su più fronti, permettendo così di coniugare la necessità di una rappresentazione esplicita, normalizzata e condivisa del contenuto, con l’esigenza derivante dal bisogno incessante di personalizzare questo contenuto, secondo prospettive soggettive condizionate dal contesto e dal punto di vista dell’utente. Alcune delle specializzazioni apportate durante la sperimentazione hanno interessato: l'uso integrato di tecniche robuste incrementali per l’annotazione automatica del testo a vari livelli di analisi linguistica; l’utilizzo di classificatori stocastici per far fronte all'inadeguatezza, dal punto di vista pratico, delle tecniche di classificazione basate sulla creazione manuale di regole simboliche; l’uso integrato di modelli standard di rappresentazione formale di conoscenza di dominio e, ancora, la messa a punto di tecniche non supervisionate di “bootstrapping della conoscenza” volte a integrare in maniera semi-automatica le attività di sviluppo e popolamento manuale del modello ontologico, per loro natura lente, ripetitive e soggette a errore.

Per maggiori approfondimenti sulle metodologie di estrazione di conoscenza dai testi basate su tecniche di NLP e sulla piattaforma software T2K², si rimanda a:

Dell'Orletta F., Venturi G., Cimino A., Montemagni S. (2014) "T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts". In Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014), 26-31 May, Reykjavik, Iceland.