TECNOLOGIE DEL LINGUAGGIO NATURALE - RADICIONI

RELAZIONE ESERCITAZIONE 2

Roger Ferrod, Simone Cullino, Davide Giosa

SOMMARIO

Il problema della disambiguazione, in inglese *Word Sense Disambiguation*, riveste particolare importanza nel settore NLP. L'obiettivo dell'esercitazione è quello di disambiguare i termini polisemici, a seconda del contesto, identificando il senso più corretto (i.e. *synset* di *WordNet*).

IMPLEMENTAZIONE

L'input dell'esercitazione consiste in due documenti testuali rappresentanti, rispettivamente, 14 frasi ambigue e un sottoinsieme del corpus *SemCor* annotato manualmente. Alla fase di analisi preliminare e pre-processing (l'xml del corpus risultava non ben formattato) segue l'invocazione dell'algoritmo di *Lesk* che, dato un termine da disambiguare e la frase che lo contiene (i.e. contesto), ritorna il senso più appropriato.
La versione semplificata dell' algoritmo di *Lesk*, utilizzata in questa esercitazione, segue un approccio *bag-of-words*, ossia rappresenta il contesto sotto forma di un insieme non ordinato di parole alle quali è stato applicato un processo di *lemmatizzazione* e eliminazione di *stop-words* e punteggiatura.
Siccome il corpus è annotato manualmente sui sensi di *WordNet*, risultata possibile calcolare l'accuratezza del metodo.

RISULTATI

L'accuratezza, in linea con le nostre aspettative, si assesta al 47%. Il metodo è stato validato sui 2 documenti di input. Nel primo documento (contenente 14 frasi in cui la parola polisemica è racchiusa tra asterischi) l'algoritmo ha disambiguato correttamente solamente 1 coppia di frasi (14% del totale). Una possibile spiegazione risiede nella limitata lunghezza delle frasi che, considerata la versione semplificata di *Lesk* e l'approccio BoW, non forniscono un contesto sufficiente a differenziare i sensi. Nel secondo documento, grazie alla maggiore quantità di informazioni, è possibile raggiungere l'accuratezza del 47%, calcolata sulla base dei 378 termini analizzati presenti in 87 frasi.