TECNOLOGIE DEL LINGUAGGIO NATURALE - RADICIONI

RELAZIONE ESERCITAZIONE 3

Roger Ferrod, Simone Cullino, Davide Giosa

SOMMARIO

La seguente esercitazione richiede di riassumere 3 documenti forniti in input secondo diverse percentuali di riduzione (e.g 10%, 20%, 30%) e secondo diverse tecniche. L'unità elementare di lavoro è la frase, ovvero non è possibile spezzare una frase; le uniche operazioni permesse sono la riorganizzazione di frasi e paragrafi.

IMPLEMENTAZIONE

È stato fornito in input anche il file contenente i vettori *Nasari* (rappresentazione *Lexicalized*).

A partire dal testo da riassumere, è possibile individuare un *topic* che lo rappresenta e utilizzarlo in seguito durante la riduzione. Il *topic* è ottenibile tramite il *Titolo* o tramite il metodo "OPP" (*Optimum Position Policy*)[1]. Estratto il *topic* desiderato, viene creato un contesto (a partire dalla rappresentazione *Bag Of Word*) collezionando i vettori *Nasari* di ogni termine del *topic*. A seconda della percentuale di riduzione, vengono mantenuti i *top-k* paragrafi con maggiore affinità con il *topic* estratto. L'affinità è stata calcolata massimizzando la *square-rooted Weighted Overlap*[2].

La tecnica "OPP" (*Optimum Position Policy*)[3] si basa sulla ricerca di frasi significative all'interno del testo in posizioni specifiche, che variano a seconda della tipologia di testo (la tipologia è stata aggiunta manualmente all'inizio del file di input). Se, ad esempio, il documento è di tipo *newspaper*, le frasi di interesse sono localizzate in: `[titolo,par2frase1,par3frase1,par4frase1,par1frase1,par2frase2,par3frase2,par4frase2,par5frase1,par1frase2,par1frase2,par6frase1]`. Nel caso specifico di *wsj - Wall Street Journal* - le frasi si possono ritrovare in: `[titolo,par1frase1,par1frase2]`.

Per ogni tecnica di estrazione del *topic*, viene creato un file di output contenente il risultato del processo di *summarization*.

RISULTATI

È possibile apprezzare una differenza, tra le due tecniche utilizzate, solamente nel terzo articolo (*The last man on the moon*). Probabilmente tale fenomeno è imputabile al fatto che la metodologia OPP attribuisce poche *features* agli articoli del *Wall Street Journal* e dunque tale tecnica è paragonabile al semplice impiego del titolo; i primi due articoli sono infatti tratti dal *Wall Street Journal* mentre il terzo è di tipo *newspaper*.

---

[1] E.H. Hovy. 1997, Identifying Topic by Position. Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP). Washington, DC.

[2] Pilehvar et. al. Align, disambiguate and walk: A unified approach for measuring semantic similarity. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics Vol. 1. 2013.

[3] E.H. Hovy. 1997, Identifying Topic by Position. Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP). Washington, DC.