Di Caro
L'orale dura circa 45 minuti. I primi 35 ti lascia parlare liberamente su tutte le esercitazioni una per una. Ti interrompe raramente, giusto per chiedere:
- Interpretami questi risultati.
- Perché hai fatto questa particolare scelta?
- Hai testato qualche altra tecnica?
Tutto molto discorsivo.
Per ottenere un ottimo voto, pretende un'ottima esposizione brillante.

Domande Di Caro

Il triangolo semiotico

E' un modello del significato in cui abbiamo:

- Concetto: rappresentazione mentale
- Rappresentazione: termine/simbolo, basata sulle convenzioni e ci permette di fare riferimento al concetto
- Referente: real word thing

AI/NLP cerca di andare verso l'alto (da rappresentazione a concetto) e anche da rappresentazione a referente (ad esempio nella computer vision)
Esiste il caso in cui non c'è la concettualizzazione (vertice) -> named identity (luoghi fisici nomi di persone organizzazioni ecc..)
Nel caso di concetti astratti invece il referente permane ma è anch'esso astratto (es giustizia pensiamo ad un caso in cui è stata fatta giustizia

Word sense disambiguation vs Word Sense Induction

Problemi di WSD:

- **Specificità**: granularità troppo fine -> troppi sensi per una parola
- **Copertura:** alcuni sottoalberi di wordnet sono poco coperti rispetto ad altri (es neologismi)
- **Soggettività**

Differenze

- Inventory vs no inventory
- Human-based vs data-based
- Grammar-based vs Usage-based (uso del linguaggio più che della struttura)
- Evaluation: semplice ma criticabile vs più complicato ma senza un ground truth.

Differenza fra Text Mining e Semantica Distribuzionale

Linguistica computazionale: Studio delle regole che governano l'espressione linguistica,

approccio top-down
Statistica (text-mining): Estrazione delle regole attraverso l'evidenza statistica, approccio bottom-up, data-driven
Semantica distribuzionale: secondo Turney coppie di parole che co-occorrono in pattern lessico-sintattici tendono ad avere una similar semantic relations.
Unisce l'approccio statistico ai task della linguistica (pos tagging, parsing..) rivisitazione in chiave linguistica delle tecniche di text-mining

Quali tipi di granularità esistono ed elencare alcuni task ad essi associati

Word - WSD WSI
Chunk - Multiword expressions
Sentence - Question answering
Discorso - Chatbot
Document - Summarization
Documents - Topic modeling

Genus differentia

Genus differentia: il genus è rappresentato dal l'iperonimo, mentre i differentia sono tutte le proprietà/attributi che caratterizzano quel concetto rispetto al suo genus (nei dizionari le definizioni vengono date in questa maniera, es. albero: è una pianta (genus) con rami foglie ecc.. (differentia))

Cos'è il topic modelling

Si tratta di un task di estrazione di topics partendo da una collections di documenti.
Un topic è una lista di termini pesati estratta dal documento.
problemi:

- non facile interpretazione del risultato (ricondurre una lista di parole ad un tema)
- A volte si estraggono topic poco utili

La LDA è una versione probabilista della LSA, assunzione: un documento è un mix di topic e ogni parola ha una probabilità di comparsa in ogni topic

Configurazioni matriciali

- **Term document** (righe docs, colonne terms) mira alla semantica del documento (ottieni un vettore per ogni doc) -> qualsiasi task legato al documento (similarity, clustering, classificazione)
- **Term context** (righe term, colonne contexts) mira alla semantica del termine, il context può essere quello che vogliamo (frase in cui compare, associazioni con altre parole, ecc..)

- **Pair Pattern** (righe coppie di termini, colonne pattern)
  Il pattern può essere più o meno quello che vuoi basta che leghi le coppie di parole a cui fai riferimento (es normalizzare il testo fra le due oppure dipendenze sintattiche)
  Cosa ci faccio? ricorda che hai un vettore per ogni coppia
    - creo cluster di coppie -> **similarità relazionale** (stai mettendo assieme coppie che hanno relazioni simili)
    - **Similarità tra pattern** (clusterizzi le colonne)
    - **Relational classification** (fai classificazione delle coppie di termini)
    - **Relational search:** tutte le X tali per cui X causa il cancro (cancro è il secondo elemento della coppia e causa è il pattern)

Semantica documentale e text tiling

Riduziona a vettore di un documento testuale (**vector space model**), vari task tra cui clustering, classification, segmentation (text tiling) e summarization.
Text tiling: individuazione del cambio di discorso
separare un testo in finestre di lunghezza fissata, si calcola la cohésion intra-gruppo dentro queste finestre e si cercano i break-point (quei punti che fanno crollare il valore di questa misura)

LSA, cos'è il concetto latente

Il **concetto latente** ha a che fare con la riduzione matriciale che permette di passare da una matrice sparsa ad una densa, analizzando le co-occorrenze dei termini.
Il concetto latente viene individuato tramite discorso indiretto (d1 e d2 usano parole diverse -> similarità pari a 0 (matrice sparsa) , ma le parole che usano sono semanticamente simili perché co-occorrono all'interno del contesto indiretto, quindi non hanno similarità nulla**)**
SVD: tecnica per riduzione vettoriale tramite combinazione lineari delle feature.

Problemi LSA:
- Non generalizza su documento mai visti (non è scalabile)
- Valori negativi di difficile interpretazione

Knowledge graph

Come puoi sfruttare la conoscenza a grafo (rdf oppure neo4j)
Gli approcci neurali sono ottimi per cogliere in maniera automatica delle relazioni tra i termini. Spesso però queste relazioni sono già state espresse in un KG -> l'unione fa la forza
- Alcuni **algoritmi** per analisi di KG:
    - Centralità di un nodo
    - Similarità tramite calcolo percorso
    - Embeddings (Node2Vec)
    - Predizione: clustering dei nodi (ottieni una comunità) oppure predizione di nuovi archi

- **Tasks**:
    - Sense disambiguation
    - Question answering -> molte info sono direttamente collegate da archi typed
    - Semantic search (ricerca non su base lessicale)
    - Sistemi di raccomandazione -> ho un nuovo nodo, stimo dei link nuovi
    - Knowledge graph completion
    - Entity resolution

Teoria di Hanks

Modelli di costruzione del significato

Per la teoria di pustejovsky non esiste una risorsa lessicale di questo tipo in quanto risulterebbe essere troppo complessa la costruzione (per quanto riguarda l'ambiguità non so..)
4 Strutture:

- **Argomentativa**
- **Eventiva**
- **Inheritance**: dove viene collocato quel termine globalmente rispetto agli altri termini (generalmente struttura tassonomica)
- **Qualia**: prorietà/attributi associabili ad una parola
    - costitutivo: legate alla fisicità
    - formale: caratteristiche distintive
    - telico: funzione per cui esiste quell'oggetto
    - agentiva: origine di un concetto

Hanks: più semplice a livello teorico ma implementabile.
L'essenza del significato è dato dal verbo. Si considera la valenza del verbo (# di argomenti del verbo) e la struttura argomentale (sogg-ogg oppure sogg-ogg-mod …)
Differenzia il significato partendo dalla valenza del verbo.
2 Arg -> 2 slot, ogni slot si compone da filler (lexical items)
Si raggruppano i filler per semantic type ottenendo varie possibili combinazioni. Ogni combinazione è un possibile significato
Problemi:

- quali semantic type e che grado di generalizzazione (questo grado dipende dal contesto)
- Abbiamo bisogno di tanti dati.