Oltre il contenuto.
Il profilo linguistico del testo

Julia Kenny

Seminario di cultura digitale

Sommario

Introduzione 2
Estrazione di conoscenza: contenuto e forma linguistica 4
Annotazione linguistica del testo 4
Estrazione di conoscenza 5
Profilo linguistico e varietà linguistiche 6
Native language identification - NLI 7
Perché 7
Come 8
Il primo shared task di NLI 10
Risultati 11
Conclusioni 14
Bibliografia e sitografia 16

Introduzione

La linguistica computazionale si occupa dell'elaborazione del linguaggio mediante l'uso di tecnologie informatiche. Nata come disciplina di frontiera ai margini sia del mondo umanistico sia di quello informatico, in poco più di cinquant'anni è riuscita a conquistare una posizione d'indiscussa centralità nel panorama scientifico internazionale e rappresenta oggi uno dei più importanti punti di contatto fra discipline umanistiche e mondo informatico.
La linguistica computazionale ha ormai raggiunto una sua maturità metodologica, nata dalla conquista di un preciso spazio di autonomia disciplinare, che si contraddistingue per un nuovo e delicato equilibrio tra lingua e computer. Dalla sua nascita alla fine degli anni '50, e dalla sua configurazione come disciplina autonoma, ha subito una crescita esponenziale in diverse direzioni arrivando ad attingere contributi da aree quali la linguistica, la psicologia, la teoria dell'informazione, la matematica, la statistica e, naturalmente, l'informatica.
La crescita di cui questa disciplina è stata protagonista è testimoniata anche dalle numerose iniziative imprenditoriali nel settore delle tecnologie della lingua, che attestano l'impatto applicativo di una disciplina matura, ormai pronta per uscire dal circoscritto ambito accademico.

Le moderne tecnologie della lingua permettono ai sistemi informatici di accedere ai contenuti di un testo in un modo nuovo. L'applicazione di tecnologie linguistiche all'analisi dell'informazione richiede una considerevole allocazione di risorse per lo sviluppo e per il processing, ma permette di aumentare la quantità e la qualità dell'informazione che è possibile estrarre da un testo, vantaggi di fondamentale importanza in anni come questi, in cui il trattamento dell'informazione è divenuto uno degli aspetti centrali della vita quotidiana di tutti noi.
Gli sviluppi più recenti nel campo del Trattamento Automatico della Lingua (TAL), o Natural Language Processing (NLP), hanno creato soluzioni tecnologiche dalle enormi potenzialità, capaci di migliorare la ricerca nei documenti testuali e in grado di gestire in modo intelligente l'informazione in essi contenuta, così da rispondere alla continua necessità di accedere a grandi quantità di contenuti digitali semi-strutturati o non strutturati.

Il problema di come acquisire e gestire la conoscenza depositata nei documenti testuali dipende dal suo essere codificata all'interno di una rete di strutture e relazioni sia grammaticali sia lessicali che costituiscono la natura stessa della comunicazione linguistica. Attraverso l'analisi automatica del testo, gli strumenti di NLP indagano la rete del linguaggio naturale per estrarre e rendere espliciti quei nuclei di conoscenza che possono soddisfare i bisogni informativi degli utenti. Dotando il computer di capacità avanzate di elaborazione del linguaggio, diventa possibile costruire automaticamente rappresentazioni del contenuto dei documenti. Queste rappresentazioni permettono di potenziare la ricerca di documenti, l'estrazione di informazione rilevante, l'acquisizione dinamica di nuovi elementi di conoscenza relativi a un certo dominio, ecc., migliorando così i processi di elaborazione e di condivisione delle conoscenze e rendendo possibile la creazione di modelli e di strumenti per il trattamento della lingua utilizzabili anche in contesti operativi reali.

Al giorno d'oggi, la ricerca nel campo della linguistica computazionale e del NLP riguarda un'ampia gamma di campi, sia teorici sia applicativi. In questo elaborato l'attenzione sarà focalizzata su uno specifico campo di utilizzo di tecniche di linguistica computazionale: la relazione si concentrerà su uno dei campi di ricerca più recenti, quello che indaga un documento testuale cercando di comprenderne non il contenuto, ma la forma.

Partendo dal seminario "Oltre il contenuto: tecnologie linguistico-computazionali per l'analisi della struttura linguistica del testo. Cosa, come, perché" si chiarisce come le tecnologie del linguaggio possano essere utilizzate per estrarre conoscenza di tipo differente: la conoscenza di dominio e la conoscenza linguistica.
L'analisi di quest'ultimo tipo di conoscenza, e quindi della forma linguistica di un testo, ha un forte potenziale innovativo in diversi settori applicativi. Si cercherà pertanto di approfondire il tema del monitoraggio del profilo linguistico e di analizzare uno dei campi in cui esso è impiegato: il riconoscimento automatico della lingua madre di testi scritti in una seconda lingua (L2) da scriventi con varie lingue madri (L1) – *Native Language Identification* (NLI).
L'approfondimento di questo argomento, brevemente presentato durante il seminario, ci permette di comprendere a pieno il reale potenziale che di un profilo linguistico estratto con le moderne tecniche di NLP e combinato all'uso dei più recenti algoritmi di *machine learning* (ML) e *data miningI* (DM).
Allo stesso tempo, è importante evidenziare come questa metodologia (profilo linguistico + algoritmi di ML) sia la stessa utilizzata per risolvere altre problematiche, ad esempio l'identificazione del genere testuale o dell'autore. Dunque, la comprensione della risoluzione di una problematica specifica come quella del NLI permette di capire meglio anche come vengono affrontate tematiche simili.

Estrazione di conoscenza: contenuto e forma linguistica

Nel seminario di cultura digitale dell'11 dicembre 2013 dal titolo "Oltre il contenuto: tecnologie linguistico-computazionali per l'analisi della struttura linguistica del testo. Cosa, come, perché" Dominique Brunato, Felice Dell'Orletta e Giulia Venturi hanno spiegato come le tecnologie linguistico-computazionali vengano oggi utilizzate per accedere al contenuto informativo di un testo, ma anche per andare oltre ad esso e accedere alla struttura linguistica di quel testo. Quest'ultima capacità - l'accesso alla struttura del testo - è un elemento di grande interesse poiché permette di ottenere un nuovo tipo di conoscenza: la conoscenza linguistica.

L'immagine seguente riassume bene buona parte dei concetti presentati durante il seminario.

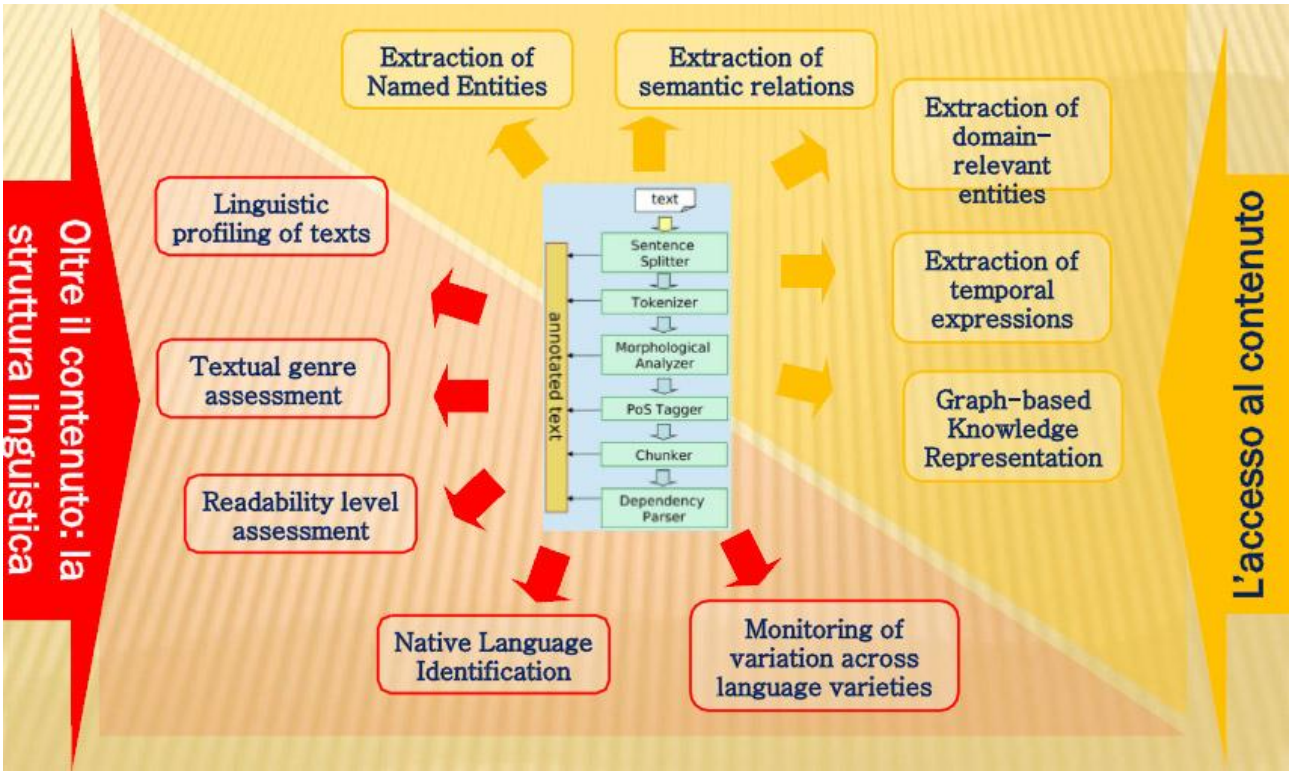

Figura 1 - Estrazione di conoscenza da testi.

In primo luogo, l'immagine evidenzia come le operazioni di accesso alla conoscenza (sia di dominio sia linguistica) non siano eseguite direttamente sul testo scritto in linguaggio naturale, ma sul testo elaborato e annotato linguisticamente.

Annotazione linguistica del testo

L'annotazione linguistica di un testo avviene secondo convenzioni e regole che dipendono dalle specifiche necessità del caso, ma i cui passaggi principali sono fondamentalmente sempre gli stessi. Infatti, l'annotazione linguistica è un processo incrementale, che avviene mediante analisi linguistiche aventi un livello di complessità di volta in volta crescente, i cui punti fondamentali sono:

1. Sentence Splitter → il testo viene segmentato in frasi;

2. Tokenizzatore → la sequenza di caratteri che costituisce una frase viene segmentata in unità minime di analisi, i cosiddetti token;
3. Analizzatore morfologico → assegna a ciascun token le possibili categorie grammaticali, spesso integrate da ulteriori specificazioni morfologiche come il genere, il numero o la persona (ad esempio, il token *danno* può essere interpretato sia come "sostantivo maschile singolare" sia come "verbo *dare* declinato alla terza persona plurale indicativo presente");
4. PoS Tagger → seleziona la corretta interpretazione morfologica del token in quel contesto specifico, in altre parole la corretta *part of speach* - POS (ad esempio, nella frase "il danno subito è stato elevato" il token "*danno*" viene indicato come sostantivo e non come verbo);
5. Chunker → aggrega i token in "chunk" sintattici (ad esempio: chunk nominale, chunk verbale, chunk preposizionale, ecc.);
6. Analizzatore sintattico → identifica le relazioni sintattiche fra i token, non lavora più a livello di parola, ma cerca di capire in che modo le parole si correlano le une alle altre. Questo tipo di annotazione risente fortemente dei diversi approcci teorici alla sintassi, che si correlano ai differenti modi di rappresentare la struttura di una frase (*a costituenti* vs. *a dipendenze*).

Il testo arricchito con informazioni linguistiche diventa il punto di partenza per ulteriori elaborazioni automatiche.

Estrazione di conoscenza

Il triangolo giallo della figura 1 rappresenta i compiti di estrazione di **conoscenza di dominio** da collezioni documentali. Rientrano in questa categoria compiti come:

- l'estrazione ed organizzazione di terminologia di dominio,
- l'annotazione semantica di entità nominate e di entità rilevanti per uno specifico dominio,
- l'estrazione di relazioni tra le entità estratte.

In tutti questi compiti si fa leva sulla semantica del documento piuttosto che sul modo in cui questa semantica è espressa. Questi ambiti di ricerca sono molto importanti e negli anni sono stati ampliamente sviluppati; le tecnologie relative a questi campi sono ormai mature.

All'opposto, nel triangolo rosa sono rappresentati i campi che riguardano l'acquisizione di **conoscenza linguistica** da corpora. L'informazione che viene estratta da queste operazioni non è più un'informazione relativa alla conoscenza di dominio, bensì un'informazione relativa all'uso della lingua. Questo è un campo di indagine più recente e fortemente innovativo che include importanti compiti di ricerca, ma che ha anche un forte impatto applicativo.
Infatti, attraverso l'analisi del profilo linguistico di un testo è possibile:

- classificare il genere testuale di un documento,
- riconoscere la lingua madre di un apprendente che scrive in una determinata seconda lingua,
- attribuire un testo a un autore,
- identificare plagi,
- misurare la leggibilità di un testo,

- semplificare un testo,
- valutare le competenze linguistiche di apprendenti la lingua (L1 o L2) nel contesto didattico.

Profilo linguistico e varietà linguistiche

Come già evidenziato in precedenza, il punto di partenza per accedere alla struttura linguistica del testo è quello di ricostruirne il profilo linguistico, in altre parole, di estrarre e monitorare le caratteristiche linguistiche di tale testo.

Usando corpora di ampie dimensioni e strumenti di analisi linguistica automatica del testo è oggi possibile condurre un monitoraggio linguistico ad ampio spettro, che coinvolge una varia ed estesa tipologia di parametri riguardanti i diversi livelli di descrizione linguistica.
Si tratta di un obiettivo ormai perseguibile perché le tecnologie linguistico-computazionali cominciano a essere mature per essere sfruttate in contesti applicativi di monitoraggio linguistico; infatti, le analisi svolte con tali tecnologie hanno raggiunto un livello di accuratezza che, seppur decrescente attraverso i diversi livelli di annotazione linguistica, è sempre più che accettabile.
Attualmente, dunque, il monitoraggio linguistico può essere condotto in relazione a corpora di vaste dimensioni e basarsi su informazioni della struttura sintattica che fino ad ora sembravano essere inaccessibili, se non attraverso un attento lavoro manuale (che per sua natura non poteva che essere limitato a porzioni di testo e circoscritto a un ristretto repertorio di tratti linguistici appositamente selezionati).

Ad oggi, il repertorio di tratti che è possibile estrarre da un testo in modo automatico sfruttando i moderni strumenti di elaborazione del linguaggio, è notevole.
Seguono esempi di caratteristiche rilevanti per i principali livelli di annotazione linguistica.

- A livello di profilo di base: il numero totale di periodi in cui si articola il testo, il numero totale di parole, la lunghezza media dei periodi (in token), la lunghezza media delle parole (in caratteri).
- A livello di profilo lessicale:
    - indici di ricchezza lessicale come il rapporto fra parole tipo/unità (TTR) o la densità lessicale, cioè il rapporto delle parole piene (ovvero portatrici di significato: nomi, aggettivi, verbi e avverbi) rispetto al totale delle occorrenze di parola all'interno del testo;
    - tipologia del vocabolario usato: percentuale di parole tipo appartenenti al vocabolario di base (VdB)[1], ripartizione delle parole rispetto ai repertori d'uso del VdB (vocabolario fondamentale, ad alto uso o ad alta disponibilità).
- A livello di profilo morfosintattico: la distribuzione delle varie POS. Infatti, la diversa distribuzione delle POS riflette differenze e somiglianze tra generi testuali, così come tra la

[1] Dizionario di riferimento: Grande Dizionario Italiano dell'uso (GRADIT, De Mauro, 2000).

lingua scritta e quella parlata (ad es. nei testi giuridici sono contenuti pochi verbi, e questa è una caratteristica discriminante dei testi di questo tipo).

- A livello di profilo sintattico: l'articolazione interna del periodo (numero medio di proposizioni per periodo, rapporto fra proposizioni principali e subordinate, ecc.), l'articolazione interna della proposizione (numero medio di parole per proposizione, numero medio di dipendenti per testa verbale, ecc.) la profondità dall'albero sintattico, ecc.

Studi sulla variazione linguistica hanno dimostrato che le varietà linguistiche (o i *registri*, secondo la denominazione degli studi di prospettiva funzionale) si differenziano per una loro propria distribuzione delle caratteristiche linguistiche. Quello che differenzia una varietà dall'altra non è l'occorrenza o meno di determinate caratteristiche all'interno del testo, ma la distribuzione di *tutte* le sue caratteristiche linguistiche. Questa distribuzione di caratteristiche, appunto il 'profilo linguistico del testo', ne caratterizza la specifica varietà e, inevitabilmente, risulta simile in testi di una medesima varietà e differente in testi di varietà diverse.

Concludendo, è oggi possibile ricostruire il profilo di generi testuali, varietà di lingua o sottolinguaggi, per poi monitorarli in modo automatico sulla base di parametri linguistici con lo scopo di descriverne la struttura linguistica e riuscire a comprenderne le differenze strutturali. Analizzando testi differenti (appartenenti a generi diversi oppure scritti da autori differenti o da apprendenti con diverse lingue madri), questo permette di riconoscere le differenze linguistiche presenti nei testi con l'obiettivo finale di eseguire compiti quali il riconoscimento del genere testuale, dell'autore, della lingua madre, ecc.
Questi compiti condividono metodologie comuni per la risoluzione di problemi diversi (ma per certi aspetti simili). Si è dunque scelto di approfondirne uno: il campo riguardante l'identificazione della lingua madre.

Native language identification - NLI

Lo scopo di un task di *Native Language Identification* (NLI) è quello di individuare in modo automatico la lingua madre (L1) di uno scrivente a partire da testi scritti in una seconda lingua (L2).

Perché

Negli ultimi anni, quest'area di ricerca ha registrato un picco di interesse, anche grazie alle forti implicazioni e livello pratico di questo campo di ricerca.
Applicazioni reali legate a compiti di NLI possono riguardare sia contesti educativi che non. In primo luogo, l'identificazione della lingua madre può essere utile in contesti educativi perché può aiutare a fornire un feedback più accurato nella correzione degli errori di testi di apprendenti: è, infatti, ben noto che apprendenti con L1 diverse commettono errori differenti, influenzati dalla loro lingua madre. Un sistema in grado di individuare la L1 di uno scrivente può comprendere meglio le motivazioni dell'errore, eseguire correzioni mirate e fornire un feedback su misura.

Inoltre, la lingua madre è una caratteristica importante per il *profiling* dell'autore di un testo ed è quindi utile in tutti i numerosi ambiti che richiedo l'identificazione dell'autore. I campi in cui l'identificazione dell'autore è richiesta sono molto vari e spesso si allontanano dall'analisi letteraria o linguistica di un testo arrivando ad abbracciare una serie illimitata di altri campi, ad esempio: la sicurezza, le ricerche di marketing o il targeting per la pubblicità. Un caso curioso e interessante di utilizzo di profiling è, ad esempio, quello della linguistica forense: la linguistica forense si occupa di studiare il ruolo, la forma e il valore di prova che il linguaggio può avere all'interno di un processo penale; a questo fine, il profiling di un parlante assume un ruolo di primaria importanza.
Infine, ricerche di NLI sono strettamente correlate ad altri campi di ricerca linguistica. Ricerche come quelle di NLI si basano su studi empirici effettuati sfruttando le caratteristiche linguistiche estratte da grosse quantità di dati e, a volte, mettono in luce aspetti importanti non ancora individuati dagli studi 'più tradizionali'. Ad esempio, ricerche di NLI possono avere influenza su studi di *acquisizione delle seconde lingue*, in particolare nello studio dei transfer linguistici che avvengono fra una lingua madre e una seconda lingua.

Come

L'identificazione della lingua madre è un campo di ricerca piuttosto recente, le prime pubblicazioni a riguardo sono quelle di Tomokiyo e Jones (2001), Jarvis et al. (2004), and Koppel et al. (2005), ma negli ultimi anni il numero di pubblicazioni in materia di NLI è notevolmente aumentato. Il crescente interesse per l'argomento ha portato a numerosi paper, progetti di ricerca e tesi di laurea o dottorato, la maggior parte dei quali si è, ovviamente, concentrata sull'identificazione della lingua madre di apprendenti l'inglese come L2.

Generalmente, il compito di NLI viene affrontato come un problema di classificazione in cui l'insieme delle lingue madri è conosciuto a priori. A partire dai dati riguardanti testi scritti da appartenenti con lingua madre appartenente all'insieme di L1 da classificare, viene addestrato un modello.
Dato un set di caratteristiche e un corpus di addestramento, il classificatore valuta la distribuzione dei tratti all'interno del campione di addestramento per ricavarne un modello matematico che formalizza il contributo di ciascun tratto (o insieme di tratti) rispetto al compito in questione. Il modello è poi applicato a esempi sconosciuti per assegnare loro la classe più probabile dato quel modello e l'insieme di tratti pertinenti.

Nonostante la crescente quantità di lavori in materia, i risultati delle ricerche svolte sono spesso difficili da comparare per motivi quali il diverso pre-processing operato sui dati, la differente suddivisione dei dati in *training set* ed *evaluation set* oppure la diversità degli insiemi di L1 analizzati.
Inoltre, le ricerche di NLI necessitano di avere a disposizione un corpus contenente testi di apprendenti una determinata L2 (tipologia di corpus più difficile da reperire rispetto ad altre tipologie di corpus) e la scarsa disponibilità di corpora di questo tipo ha portato molte ricerche a sfruttare il corpus ICLE, fino a qualche anno fa uno dei pochi corpora disponibili contenenti testi di

inglese come L2, ma comunque non adatto a compiti di NLI. ICLE contiene qualche centinaia di temi scritti da studenti del college apprendenti l'inglese, ma è un corpus troppo piccolo per svolgere analisi di tipo statistico. Non è possibile dire se risultati ottenuti con questo corpus siano scalabili su dataset più grandi o su dataset di dominio differente.

Avere uno spazio per mettere a confronto tecniche adoperate e risultati ottenuti è fondamentale per aiutare il progresso di un campo di ricerca, che può evolversi solo mediante il confronto degli approcci più tradizionali con quelli più innovati e attraverso una conoscenza condivisa di cosa effettivamente funziona e cosa, invece, è meglio evitare. Una valutazione controllata, oggettiva e gestita da un unico organismo, consente di valutare realmente le prestazioni delle differenti soluzioni al problema proposte dai vari studiosi, consentendo alla comunità una reale definizione dello "stato dell'arte", delle potenzialità e delle problematicità dei vari approcci. Con questo obiettivo vengono spesso organizzate campagne di valutazione relative a un preciso ambito di ricerca.
In quest'ottica, il primo shared task di NLI[2] rappresenta un primo tentativo di valutazione delle ricerche nel campo dell'individuazione automatica della lingua madre a partire da testi scritti.

Il primo shared task di NLI si è tenuto il 13 e 14 giugno 2013 ad Atlanta, negli Stai Uniti, all'interno del workshop sull'utilizzo di NLP per la creazione di applicazioni educative, "*The 8th Workshop on Innovative Use of NLP for Building Educational Applications*"[3]. È importante sottolineare come il successo di questo workshop, arrivato nel 2015 alla sua decima edizione, testimonia a livello internazionale l'affermarsi dell'uso di tecnologie del linguaggio per lo studio dei processi di apprendimento. Infatti, l'uso di tecnologie del linguaggio per lo studio dei processi di apprendimento è sempre di più al centro di ricerche interdisciplinari che mirano a mettere in luce come metodi e strumenti di annotazione linguistica automatica e di estrazione della conoscenza siano oggi maturi per essere usati anche in applicazioni educative e scolastiche, come la costruzione di sistemi intelligenti di supporto all'insegnamento delle lingue (*Intelligent Computer–Assisted Language Learning systems* - ICALL).

Un'analisi più approfondita di come è stato organizzato lo shared task di NLI e dei contributi dei partecipanti ci permette di comprendere nel dettaglio in cosa consiste un'analisi di riconoscimento della lingua madre e di comprendere il ruolo che le caratteristiche linguistiche hanno all'interno di questa analisi.

[2] NLI Shared task 2013, sito web: https://sites.google.com/site/nlisharedtask2013/home
[3] The 8th Workshop on Innovative Use of NLP for Building Educational Applications, sito web: http://www.cs.rochester.edu/~tetreaul/naacl-bea8.html

Il primo shared task di NLI

L'obiettivo del primo shared task di NLI è quello di facilitare il confronto delle metodologie di ricerca. Il problema della scarsa possibilità di comparazione riscontrato nelle analisi fin ora svolte viene risolto in questo task condiviso

- sfruttando un corpus ampio e appositamente realizzato per compiti come il NLI
- fornendo un set comune di L1 da analizzare e standard di valutazione che ogni partecipante alla competizione deve adoperare.

Il dataset utilizzato per lo shared task è ottenuto a partire dal corpus TOEFL11. Esso è formato da saggi di apprendenti l'inglese come L2 realizzati per il test di ammissione all'università "*Test of English as a Foreign Language*" (TOEFL®).
I corpora di testi realizzati come L2 sono da tempo uno strumento essenziale per lo studio dell'apprendimento delle lingue, e il corpus TOEFL è stato realizzato appositamente per permettere le più moderne analisi automatiche come l'identificazione della lingua madre, l'individuazione e correzione degli errori grammaticali e la valutazione automatica degli elaborati. Viste le sue ampie dimensioni e il bilanciamento nella sua composizione interna, il corpus risulta ad oggi il più indicato per un compito di NLI di scritti in lingua inglese di parlanti non nativi.
Il dataset contiene 1100 scritti per L1 ed è stato composto nel modo il più equilibrato possibile cercando di bilanciare:

- le 11 lingue madri degli scriventi - arabo (ARA), cinese (CHI), francese (FRE), tedesco (GER), hindi (HIN), italiano (ITA), giapponese (JAP), coreano (KOR), spagnolo (SPA), telugu (TEL), e turco (TUR),
- gli 8 argomenti assegnati nel test come tema dell'elaborato,
- i 3 livelli di competenza degli scriventi (alto/medio/basso).

Task come il NLI richiedono sempre che un modello di classificazione sia creato a partire dai dati e poi testato su dati sconosciuti; per questo motivo i dati del corpus da utilizzare devono essere suddivisi e impiegati in parte per sviluppare i modelli e in parte per testarli.
Per lo shared task il corpus è stato quindi suddiviso in 3 set:

- *training set* o set di addestramento (TOEFL11-TRAIN) – 900 testi per ogni L1,
- *development set* o set di sviluppo (TOEFL11-DEV) – 100 testi per ogni L1,
- test set (TOEFL11-TEST) – 100 testi per ogni L1.

Infine, lo shared task è stato suddiviso in tre sotto-task che si differenziano per i corpora che è possibile utilizzare nella creazione del modello: il task principale (*Closed-training*) richiedeva di creare un modello per l'individuazione delle 11 lingue madri basandosi solo su TOEFL11-TRAIN ed, eventualmente, TOEFL11-DEV; mentre gli altri due task richiedevano di usare una combinazione di questi corpora con corpora esterni oppure esclusivamente corpora esterni.
Per ogni sotto-task vengono utilizzati corpora di addestramento differenti, ma i modelli sviluppati sono tutti testati sul medesimo test set, TOEFL11-TEST, così da rendere possibile il confronto dei risultati.

Ventinove team hanno partecipato ad almeno uno dei task dello shared task. Di questi, ventiquattro hanno scritto un report relativo al lavoro svolto. È utile consultare i report per comprendere come i vari team hanno affrontato il problema e sviluppato una soluzione.

Risultati

In generale, tutti i team si sono posti l'obiettivo di ottimizzare il problema di NLI, cioè, di migliorare il più possibile l'accuratezza dei risultati di classificazione ottenibili con un determinato modello. L'accuratezza è una metrica relativa alla qualità del modello di classificazione che si ottiene osservando per ogni classe (in questo caso per ogni L1) quanti testi sono stati correttamente riconosciuti come appartenenti a quella classe, rispetto a quelli che realmente appartenevano a tale classe.

Cioè che distingue la metodologia di un gruppo da quella di un altro sono le caratteristiche adoperate per creare il modello e gli algoritmi di apprendimento automatico utilizzati per realizzarlo. Le tecniche di *machine learning* utilizzate costituiscono un argomento di analisi molto interessante, ma una loro trattazione esaustiva richiede conoscenze approfondite che esulano dai concetti che questo elaborato vuole presentare. Invece, quello che è importante approfondire ai fini di questo elaborato sono le caratteristiche utilizzate per la creazione del modello.
Analizzando i report, per quanto riguarda le tecniche di *machine learning* utilizzate, ci limitiamo qui a sottolineare che, nonostante esistano molti algoritmi di apprendimento automatico da adoperare per problemi di classificazione, la maggior parte dei team ha scelto di usare il *Support Vector Machine* - SVM, così come avevano già fatto i gruppi che avevano scritto i precedenti lavori sull'argomento. Alcuni team hanno provato a usare anche metodi *ensamble,* cioè classificatori multipli che predicono la classe risultante non sulla base di un unico modello, ma sulla base di combinazioni di più modelli. Infine, alcuni gruppi hanno provato algoritmi alternativi a quelli 'tradizionali' come il *Maximum Entropy* e il *K-Nearest Neighbors*.

Per quanto riguarda le caratteristiche usate, le più comuni sono state: parole, caratteri e n-grammi di parti del discorso (part of speach – POS). La maggior parte delle squadre che ha utilizzato n-grammi ha usato unigrammi, bigrammi o trigrammi, scelta in linea con quelle effettuate nella letteratura precedente. Ma ci sono stati anche team che hanno usato n-grammi di ordine superiore; addirittura, quattro delle prime cinque squadre che si sono classificate come migliori per livello di accuratezza hanno usato almeno 4-grammi, mentre altri sono arrivati a usare anche 7-grammi o 9-grammi.
Inoltre, sei team hanno usato anche caratteristiche sintattiche, principalmente alberi di dipendenze; di queste però, solo due sono risultate fra le migliori dieci per accuratezza. Tre team hanno usato feature relative allo spelling.
Infine, sono state usate anche caratteristiche nuove rispetto a quelle 'tradizionali' o comunque già sperimentate in precedenza: alcuni team hanno evitato caratteristiche classiche come le parole o le POS a favore di metriche innovative basate su frequenze relative di POS e lemmi di parola oppure a favore dell'utilizzo di alcune misure di complessità del testo.

In particolare, nel contributo di Andrea Cimino, Felice Dell'Orletta, Giulia Venturi e Simonetta Montemagni "Linguistic Profiling based on General–purpose Features and Native Language Identification" viene sottolineato come sia possibile risolvere questo compito di classificazione delle lingue madri utilizzando come feature per il modello di classificazione esclusivamente caratteristiche linguistiche, cioè elementi che descrivono la forma degli scritti. L'idea è quella di utilizzare non feature selezionate ad hoc fra quelle considerate più rilavanti negli studi precedenti, ma di monitorare il profilo linguistico degli scritti di apprendenti con le varie L1 per selezionare le più discriminanti direttamente mediante l'osservazione dei dati. Come già sottolineato più volte, il profilo linguistico permette di identificare gruppi di testi simili, almeno rispetto alle caratteristiche descritte nel profilo.

I profili linguistici degli undici sotto-corpora (contenenti ciascuno tutti i testi scritti da parlanti aventi una delle undici L1 prese in considerazione) risultano simili

- sia quando ogni sotto-corpora è stato creando utilizzando i dati provenienti dal corpus di addestramento,
- sia quando è stato utilizzato solo quello di sviluppo,
- sia quando sono stati usati entrambi i corpora in maniera combinata.

Questa può essere considerata una prova sia dell'affidabilità dell'approccio basato sull'analisi del profilo linguistico sia della rilevanza delle caratteristiche linguistiche per compiti quali un' analisi di NLI.

Vengono ora ripresi alcuni grafici dal report di questo team per mostrare le differenze fra le caratteristiche dei vari sotto-corpora.

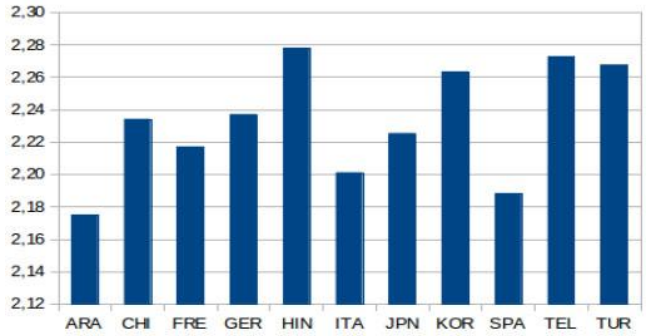

Grafico 1 - Lunghezza media delle parole

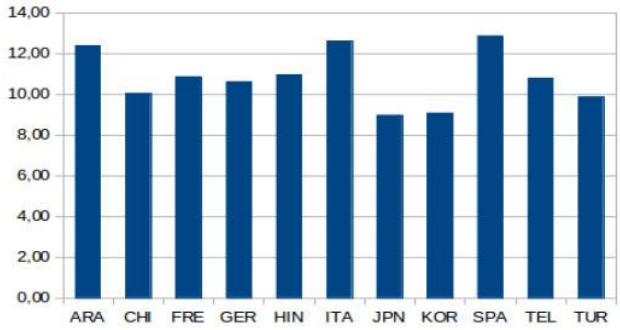

Grafico 2 - Lunghezza media dei periodi.

Partendo dalle caratteristiche di base è possibile osservare che sia la lunghezza media delle parole sia la lunghezza media dei periodi varia significativamente fra le varie L1.
In particolare, gli scritti di apprendenti con L1 arabo e quelli di apprendenti con L1 spagnolo sono caratterizzati dal fatto di contenere parole in media più corte e periodi in media più lunghi, mentre le parole più lunghe si trovano nei testi di hindi e telugu e i periodi più brevi si trovano in quelli di giapponesi. Apprendenti con altre lingue madri, come il cinese, hanno una lunghezza media di parole e di periodi in linea con la distribuzione generale, che non li distingue in modo particolare da altri scriventi, almeno per quanto riguarda queste caratteristiche.

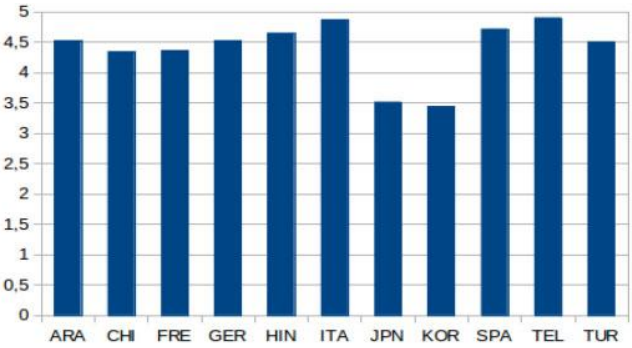

Grafico 3 - Distribuzione degli aggettivi.

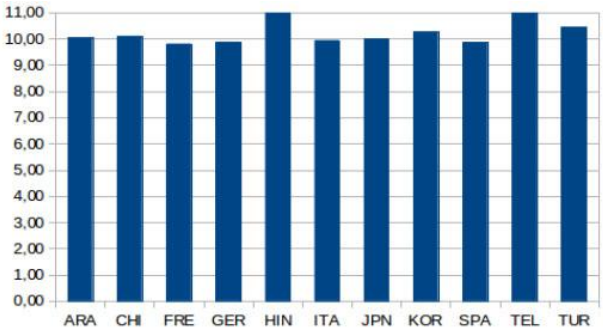

Grafico 4 - Distribuzione dei nomi.

La distribuzione delle POS è stata spesso utilizzata per individuare differenze fra varietà linguistiche. Osservando la distribuzione di nomi e aggettivi, due caratteristiche già utilizzate in molti studi di NLI, si osserva che i vari sotto-corpora hanno una distribuzione simile. Si discostano da questi in modo particolare solo i testi dei giapponesi e quelli dei coreani, che mostrano la percentuale più bassa di aggettivi, e quelli di hindi e telugu, che sono caratterizzati dalla maggiore percentuale di nomi.

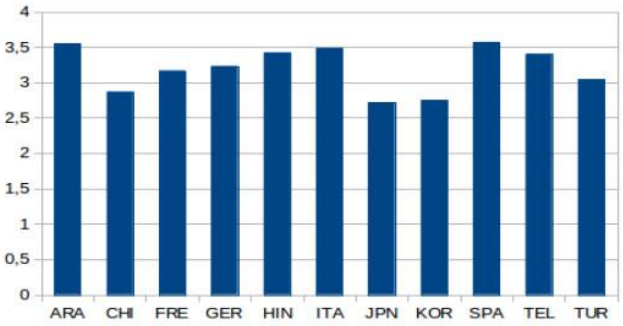

Grafico 5 - Altezza media degli alberi di dipendenza.

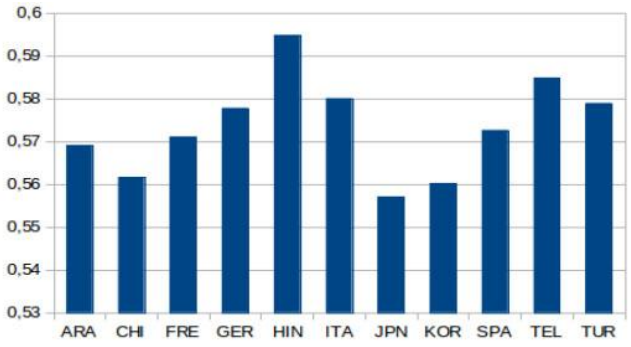

Grafico 6 - Profondità media delle catene di subordinazione.

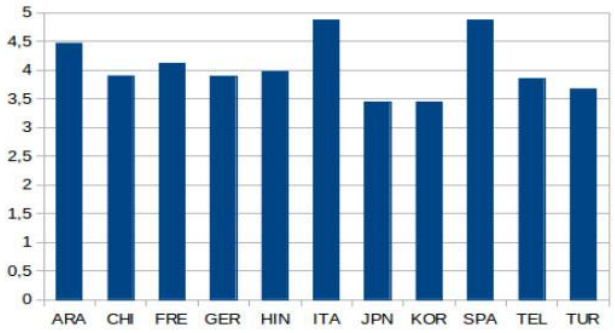

Grafico 7 - Media delle lunghezze massime delle relazioni di dipendenza.

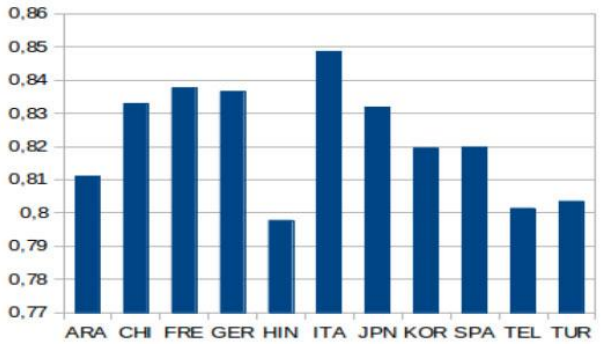

Grafico 8 - Arità media delle teste verbali.

Arrivando ad analizzare le caratteristiche sintattiche, si osserva che i testi dei giapponesi e dei coreani si distinguono fortemente da quelli degli altri apprendenti per una differente distribuzione dei tratti linguistici. Nello specifico, hanno 'alberi di dipendenza' meno profondi di tutti, le 'catene di subordinazione' meno profonde e le più corte 'lunghezze massime di relazioni di dipendenza governate da una testa nominale'.

Dall'altro lato, gli 'alberi di dipendenza' più profondi si registrano nei testi scritti da spagnoli e arabi, mentre le 'catene di subordinazione' più profonde si ritrovano negli scritti di hindi e telugu e le più lunghe 'lunghezze massime di relazioni di dipendenza governate da una testa nominale' si trovano nei testi di italiani e spagnoli.
Infine, mentre gli scritti di telugu e hindi presentano un valore molto basso di 'arità media delle teste verbali', quelli degli italiani presentano il valore più alto.

Già analizzando pochissime caratteristiche è possibile comprendere come alcuni sotto-corpara si distinguano dagli altri per una particolare distribuzione di tali caratteristiche. Tuttavia, analizzando poche caratteristiche i corpora la cui distribuzione delle caratteristiche analizzate segue quella media degli altri corpora non vengono adeguatamente caratterizzati. È per questo motivo che per la creazione del modello di classificazione vengono utilizzate molte più feature e fra queste vengono selezionate in modo automatico solo quelle più rilevanti. La selezione automatica effettuata con algoritmi appositi garantisce che siano selezionate proprio le caratteristiche che permettono di creare il modello più accurato possibile.

Una cosa che è interessante notare, anche per riprendere il discorso sul contributo che analisi di questo tipo possono apportare ad altri tipi di studi linguistici, è che i profili linguistici dei sotto-corpora mostrano tendenze simili fra i testi di scriventi con diverse L1 a differenti livelli di analisi linguistica. Per esempio, si può osservare che scritti di giapponesi, coreani, italiani e spagnoli – L1 che appartengono a due famiglie linguistiche molto differenti - mostrano una simile distribuzione di alcune caratteristiche. Altre similarità sono state riscontrate nei sotto-corpora scritti da hindi e dei telugu, ancora due lingue appartenenti a famiglie linguistiche diverse.
Le similarità nelle caratteristiche di scritti di persone che hanno lingue madri appartenenti a famiglie linguistiche diverse, spesso anche molto lontane fra loro, è un fattore curioso, che può essere spunto per ricerche linguistiche 'più tradizionali'.

Conclusioni

Le scienze umane sono da tempo entrare nell'era digitale e la linguistica computazionale, con la sua forte interdisciplinarietà e il suo forte impatto innovativo, ne rappresenta un ottimo esempio.

Se da un lato le tecnologie linguistico-computazionali svolgono un ruolo ormai indiscusso per l'accesso al contenuto testuale, il loro utilizzo non appare ancora altrettanto rilevante quando si considera il loro contributo nella valutazione delle strutture linguistiche sottostanti al testo.
Tuttavia, negli ultimi anni analisi della struttura linguistica stanno diventando sempre più centrali nel settore del trattamento automatico del linguaggio, presentano un forte potenziale innovativo in diversi settori applicativi.

Questo contributo ha cercato di approfondire il tema del monitoraggio del profilo linguistico di un testo e di mostrarne il potenziale analizzando un preciso caso di studio: la *native language identification* - NLI.

Il problema di NLI è solitamente affrontato come un task di classificazione del testo eseguito combinando metodi di *natural language processing* (per l'estrazione delle caratteristiche) con algoritmi di apprendimento automatico (per la realizzazione del modello).

Ciò che è importante comprendere è come la stessa metodologia sia condivisa da altri task, ad esempio, il riconoscimento dell'autore, l'attribuzione di paternità di un'opera, l'identificazione del genere o la valutazione della leggibilità. Tralasciando le evidenti differenze che si riscontrano a livello di tipologia di caratteristiche linguistiche selezionate e a livello di tecniche di apprendimento automatico utilizzate, questi diversi compiti condividono un approccio comune ai problemi che affrontano: riescono a determinare la varietà di lingua, l'autore, il genere di testo o il livello di leggibilità di un testo sfruttando la distribuzione di vari tipi di caratteristiche linguistiche estratte automaticamente da testi e i moderni algoritmi di *machine learning*.

Bibliografia e sitografia

- Barbagli, Alessia, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi. 2014. *Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilita di scrittura nella scuola secondaria di primo grado*. http://clic.humnet.unipi.it/proceedings/vol1/CLICIT201415.pdf

- Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill e Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English*. http://www.ets.org/Media/Research/pdf/RR-13-24.pdf

- Brunato, Dominique, Felice Dell'Orletta e Giulia Venturi. *Oltre il contenuto: tecnologie linguistico-computazionali per l'analisi della struttura linguistica del testo. Cosa, come, perché*. Seminario di cultura digitale, 11 dicembre 2013. http://labcd.humnet.unipi.it/seminario/cultura_digitale67-32/2013/12/12/oltre-il-contenuto-tecnologie-linguistico-computazionali-per-lanalisi-della-struttura-linguistica-del-testo/

- Calzolari , Nicoletta e Alessandro Lenci. 2004. *Linguistica computazionale - strumenti e risorse per il trattamento automatico della lingua*. In "Mondo digitale", 3, pp. 56-69. http://www.mobilab.unina.it/Resources/Master%20MSTD-Mazzeo/Lenci_p._56-69.pdf

- Cimino, Andrea, Felice Dell'Orletta, Giulia Venturi e Simonetta Montemagni. 2013. *Linguistic Profiling based on General–purpose Features and Native Language Identification*. http://aclweb.org/anthology/W/W13/W13-1727.pdf

- Koppel, Moshe, Jonathan Schler e Kfir Zigdon. 2005. *Determining an Author's Native Language by Mining a Text for Errors*. http://eprints.pascal-network.org/archive/00001433/01/p342-koppel.pdf

- Montemagni, Simonetta. 2014. *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*. http://www.italianlp.it/wp-content/uploads/2014/04/montemagni_silta_submission_rif.pdf

- NLI Shared task 2013.
  Sito web: https://sites.google.com/site/nlisharedtask2013/home

- Tetreault , Joel, Daniel Blanchard e Aoife Cahill. 2013. A Report on the First Native Language Identification Shared Task. http://aclweb.org/anthology/W/W13/W13-1706.pdf

- The 8th Workshop on Innovative Use of NLP for Building Educational Applications.
  Sito web: http://www.cs.rochester.edu/~tetreaul/naacl-bea8.html