SAPIENZA
UNIVERSITÀ DI ROMA

Allineamento traduzioni Tedesco-Inglese per trasferire informazioni di coreference

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica

Corso di Laurea in Informatica

Candidato

Aurora Polifemo
Matricola 1802485

Relatore

Andrea Sterbini

Anno Accademico 2019/2020

Indice

Capitolo 1

Introduzione

Il linguaggio umano è una delle parti più variegate e complesse dell'essere umano, ma è proprio la nostra abilità di parlare e comunicare ad averci portato fin dove ci troviamo ora. Sono più di 7000 le lingue parlate nel mondo e come per un essere umano anche per un calcolatore è difficile capirle e interpretarle tutte. Tuttavia questo non significa che sia impossibile utilizzare un elaboratore per analizzarle.

Ciò che noi intendiamo come linguaggio umano, nel campo dell'informatica è definito come "lingua naturale", così indicato affinché lo si possa differenziare da quello più meccanico, il "linguaggio macchina". È in questo contesto che appare indispensabile introdurre i concetti di "Natural Language Processing" (NLP), un'ampia area tecnologica ancora in vasta crescita e sviluppo, e di Text Mining. Ma di che cosa si occupano?

L'NLP è una parte dell'informatica e dell'Intelligenza Artificiale che si occupa del linguaggio umano e il suo compito primario è elaborare la lingua naturale a partire da un calcolatore, di conseguenza quello di trasformare il testo in dati che possono essere analizzati dallo stesso calcolatore.

Mentre il Text Mining è un processo di derivazione di importanti informazioni dai testi scritti in lingua naturale.

L'NLP ha diverse applicazioni, tra le più importanti ricordiamo il sentiment analysis, il riconoscimento vocale, come Siri, Cortana e l'assistente Google, la traduzione automatica, come Google Traduttore, e l'estrazione delle informazioni. Sono vari gli step che vengono fatti quando si lavora con l'NLP (*Figura 1.1*), tra questi solo alcuni verranno utilizzati in questa ricerca; i seguenti:

- Tokenization: processo che divide una stringa in "token".
- Stemming: processo che porta ogni parola alla sua forma radice/base.
- Lemmatization: considera l'analisi morfologica della parola, simile allo stemming, in quanto raggruppa diverse forme della stessa parola considerando la forma base, ma ha output diverso.
- Part-of-Speech Tagging: processo che specifica quale parte del discorso è una parola in una determinata frase.
- Named Entity Recognition: processo di individuazione delle named entities. Una named entity è un "oggetto del mondo reale" a cui viene assegnato un nome, ad esempio una persona, un paese, un prodotto o il titolo di un libro.
- Chunking: processo che prende delle singole parti di informazioni su una frase e le raggruppa in un "chunk".

Dopo aver applicato i diversi procedimenti sui documenti di interesse, questi sono pronti per essere letti ed interpretati da una macchina.

È stato introdotto l'NLP con queste sue caratteristiche per poter spiegare al meglio ciò di cui tratterà questo elaborato, poiché è parte dello studio della lingua naturale, basandosi però esclusivamente sull'analisi di testi.

L'obiettivo principale di questo studio, descritto in dettaglio nei capitoli 4 e 5, è individuare le corrispondenze tra dei romanzi in lingua inglese e in lingua tedesca ed estrarre e trasferire delle informazioni sulle pronominal coreference.

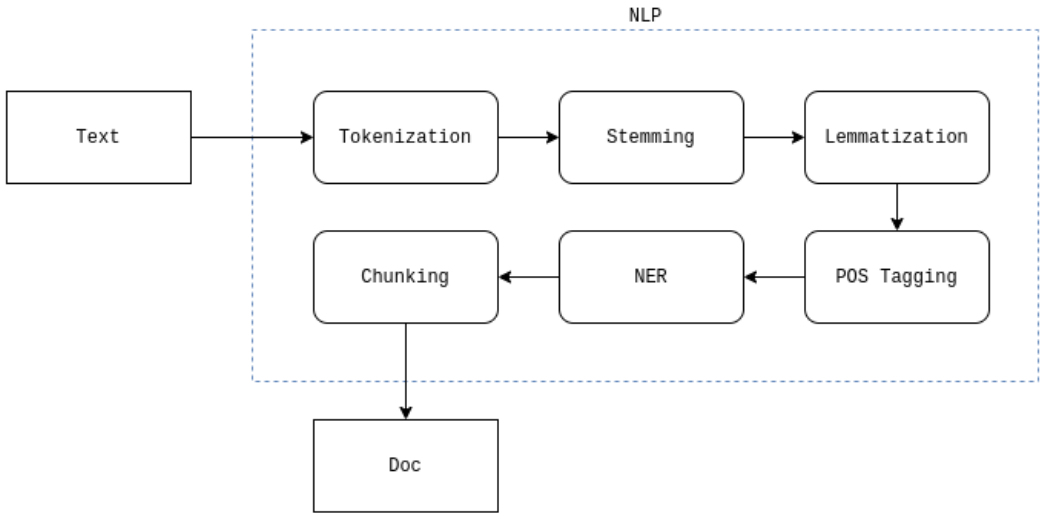


Figura 1.1. NLP pipeline

Nella prima parte verranno presi in considerazione sia dei testi originali inglesi con i rispettivi testi tradotti in lingua tedesca, sia dei testi originali tedeschi con le rispettive traduzioni inglesi. L'obiettivo è capire se, considerata una frase di un testo e il suo corrispettivo nel testo tradotto, esse risultano effettivamente essere una la traduzione dell'altra.

In sintesi, il compito principale della ricerca è mettere in corrispondenza frasi che abbiano lo stesso significato in due lingue differenti e valutare che sia stato fatto correttamente. Sono stati considerati degli approcci che variano in base a come valutare la corrispondenza e in base alla quantità di frasi considerate, come si potrà approfondire nel capitolo a ciò dedicato.

In precedenza sono state già fatte delle ricerche simili anche se non del tutto pertinenti al nostro caso, motivo per cui non si è scelto di seguire quei tipi di approcci. Tuttavia, si ritiene opportuno citarne alcuni.

Il primo tra questi è il cosiddetto modello "Bag-of-Words" utilizzato per confrontare più documenti. Il modello BoW è una metodologia per estrarre delle informazioni da un testo, raggruppando le parole tra loro e scartando tutti gli altri

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Figura 1.2. Esempio di vettori del BoW per le frasi 1.*"This movie is very scary and long."*, 2.*"This movie is not scary and is slow."*, 3.*"This movie is spooky and good."*

aspetti della frase, come la struttura e l'ordine. È di interesse sapere se occorre una nota parola e non dove è posizionata nel documento. Quello che si va a costruire è un vettore di 1 e 0, dove l'1 indica se la parola è presente nella frase e 0 se non lo è. È con questi vettori che si cerca di trovare la somiglianza tra una frase e l'altra (*Figura 1.2*).

Nonostante l'approccio utilizzato in questo elaborato derivi in parte dal modello BoW, non è stato usato questo metodo nella sua totalità in quanto, trattandosi di lingue differenti, non è possibile raggruppare parole simili che compaiono in entrambi i testi.

Un altro approccio fa riferimento allo studio che si è dedicato a riassumere più documenti trattanti lo stesso argomento, senza perdere le informazioni più importanti contenute in essi (Özateş et al.). I ricercatori ritenevano che il modello BoW non fosse abbastanza adatto per catturare la similarità sintattica e semantica tra i testi, per questo motivo hanno deciso di usare l'albero delle dipendenze.

Le dipendenze mostrano quale è la relazione di una parola con un'altra in una determinata frase, tra queste abbiamo la radice (root) dell'albero, normalmente rappresentata dal verbo della frase, che è in relazione con il soggetto (nsubj), o con l'oggetto (dobj), oppure un oggetto in relazione con un articolo (det), come si può vedere in *Figura 1.3*.

Esistono differenti tipi di relazioni di dipendenza e non tutti hanno la stessa

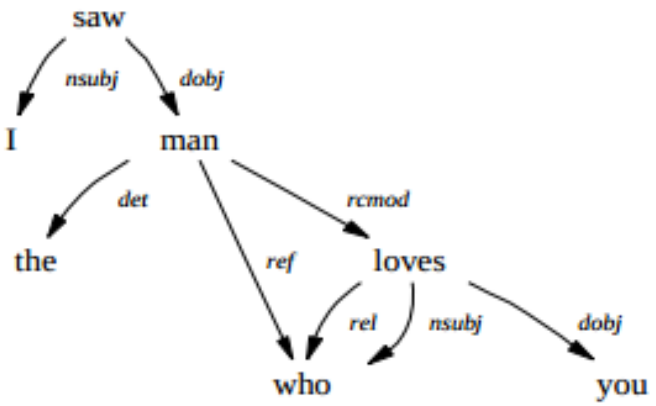


Figura 1.3. Dependency Tree della frase *"I saw the man who loves you."*

importanza. Quello di cui hanno fatto uso è, infatti, un albero in cui considerano il *tipo* di relazione che vige tra una parola dipendente, *dependent*, e una parola "radice", la *head*, questi tre elementi vengono definiti come *bigram units*. La loro idea è quella di ricavare delle informazioni sulla similarità semantica tra due frasi che hanno le stesse bigram units. Questo è conveniente quando si adoperano proposizioni nella stessa lingua, in quanto grammaticalmente seguono la stessa struttura sintattica, nel nostro caso, invece, dovendo lavorare con frasi in lingue diverse, le relazioni tra le parole e la struttura stessa della frase sono differenti.

Nella seconda parte, invece, saranno di particolare interesse gli originali tedeschi, in quanto l'intento è quello di trasferire delle informazioni, ricavabili grazie ad altri studi fatti precedentemente, dal testo in inglese al testo tedesco. L'informazione che si vuole trasferire è quella della pronominal coreference. Verranno considerati i pronomi che si riferiscono ai personaggi nel testo.

Oltre a queste due parti, abbiamo anche una terza parte, descritta nel capitolo 6, in cui verranno combinati i due lavori citati sopra, l'allineamento e il trasferimento di informazioni. In poche parole si cercherà di ricavare le informazioni dalle frasi inglesi che risulteranno essere la corretta traduzione del testo tedesco.

Capitolo 2

Strumenti e risorse utilizzate

Innanzitutto è necessario introdurre gli strumenti utilizzati per poter lavorare su dei testi in lingua naturale e per poterne estrarre diverse informazioni che verranno adoperate nel corso dell'elaborato.

2.1 BookNLP

BookNLP è una pipeline del natural language processing che si adatta a libri e altri documenti lunghi in lingua inglese ed include:

- Part-of-speech tagging
- Dependency parsing
- Named entity recognition
- Character name clustering
- Quotation speaker identification
- Pronominal coreference resolution
- Supersense tagging

I primi tre processi sono stati definiti nel capitolo 1, dunque non entreremo nel loro merito, mentre gli altri quattro verranno descritte di seguito.

La ricerca di Bamman si basa principalmente sull'analizzare dei testi, appartenenti ad un dataset di 15.099 romanzi distinti, e ricavarne delle importanti informazioni come quelle sui personaggi e sulle relazioni tra essi.

Dedurre un personaggio dal contesto non è tanto facile quanto possa sembrare, poiché non sappiamo bene cosa si intende con il termine "personaggio". È universalmente noto che sia la rappresentazione di una persona reale o immaginaria, ma esiste una forte tradizione critica che vede un personaggio più come una dimensione formale della narrazione, dunque è di interesse la funzione narrativa del personaggio più che le sue caratteristiche. È in questo ambito che si fa quindi la distinzione tra modello *referenziale* e *formalista* dei personaggi.

L'obiettivo di BookNLP è quello di fornire un metodo computazionale che possa distinguere queste due dimensioni.

Abbiamo quattro categorie di relazione di dependency tra i personaggi:

1. agent: indica le azioni di un personaggio, dunque l'atto che sta compiendo. Sono i verbi per i quali il personaggio ha una relazione di "nsubj" o "agent".

2. patient: rappresentano le azioni di altri sul personaggio. Sono i verbi per i quali il personaggio ha una relazione di "dobj" o "nsubjpass".

3. possessive: sono gli oggetti in possesso del personaggio. Sono tutte le parole che hanno una relazione "poss" con il personaggio.

4. predicative: sono caratteristiche che lo descrivono. Sono tutti gli aggettivi e nomi che hanno una relazione "nsubj" con il personaggio.

Questa distinzione rende più facile lavorare sui personaggi negli step successivi.

Il *Character Clustering* si dedica a distinguere quando un personaggio è stato solo menzionato, come nel titolo di un romanzo, oppure se è un'entità.

La *Quotation speaker identification*, come suggerisce il nome, identifica quando nel testo ricorre un discorso diretto e ne associa gli interlocutori.

La *Pronominal coreference resolution* è stata introdotta in quanto dal Character Clustering si è notato che la maggior parte dei personaggi vengono identificati in forma di pronome e non sempre con un nome proprio, infatti sui 15.099 testi, ci sono 1.673 menzioni di nomi e 4.641 sotto forma di pronomi.

Il *Supersense tagging* considera i sostantivi, i verbi, gli aggettivi e gli avverbi e ai quali associa delle informazioni aggiuntive, ricavate da WordNet, in modo da arricchire il loro significato.

In questo elaborato è stato utilizzato BookNLP in particolare per la Pronominal Coreference Resolution. Per ogni romanzo inglese preso in considerazione è stato eseguito il comando che crea tre file differenti, un file book che contiene informazioni sulle quattro categorie descritte sopra, un file token che presenta le informazioni sul singolo token, e un file html in cui si possono ritrovare le informazioni sulle coreference. Solo quest'ultimo è stato utilizzato. Dunque i testi inglesi con le coreference, che verranno utilizzati per il trasferimento di informazioni nel testo tedesco, sono stati ricavati in questo modo.

2.1.1 CoreNLP

Stanford CoreNLP fornisce una serie di strumenti di analisi del linguaggio naturale scritti in Java. Il suo input è un testo in linguaggio umano grezzo e fornisce le forme base delle parole, le loro parti del discorso, sia che si tratti di nomi di aziende, che di persone, o altro. È utilizzato per normalizzare e interpretare date, orari e quantità numeriche, per contrassegnare la struttura delle frasi in termini di frasi o dipendenze di parole e indica quali frasi nominali si riferiscono alle stesse entità.

È stato originariamente sviluppato per l'inglese, ma ora fornisce anche diversi livelli di supporto per altre lingue come l'arabo, il cinese, il francese, il tedesco e lo spagnolo.

È un framework integrato, che rende molto facile l'applicazione di una serie di strumenti di analisi del linguaggio a una parte di testo.

Stanford CoreNLP è un insieme di strumenti di elaborazione del linguaggio naturale stabili e ben testati, ampiamente utilizzati da vari gruppi nel mondo accademico, industriale e governativo.

2.2 WordNet

WordNet è un grande database lessicale di inglese. Nomi, verbi, aggettivi e avverbi sono raggruppati in insiemi di sinonimi cognitivi, i cosiddetti *synset*, ognuno dei quali esprime un concetto distinto. I synset sono interconnessi per mezzo di relazioni concettuali, semantiche e lessicali. WordNet raggruppa le parole in base al loro significato. Tuttavia, ci sono alcune importanti distinzioni. Per primo, WordNet collega non solo le forme delle parole - stringhe di lettere - ma i significati specifici delle parole. Di conseguenza, le parole che si trovano in stretta vicinanza l'una all'altra nella rete sono semanticamente disambiguate. In secondo luogo, WordNet etichetta le relazioni semantiche tra le parole.

In un primo momento è stato usato WordNet per lavorare sui testi. Procedendo con lo studio, però, si è potuto notare che i synset che ricavava dalle parole tedesche erano in numero molto inferiore rispetto a quelle inglesi, rendendo il confronto molto sbilanciato. Per questo motivo si è passati a BabelNet.

2.3 BabelNet

BabelNet è sia un dizionario enciclopedico multilingue, con copertura lessicografica ed enciclopedica dei termini, sia una rete semantica che collega concetti ed entità nominate in una rete molto ampia di relazioni semantiche, composta da circa 16 milioni di voci, chiamate "Babel Synset". Ogni babel synset rappresenta un dato

significato e contiene tutti i sinonimi che esprimono quel significato in una gamma di lingue diverse. BabelNet comprende 284 lingue ed è ottenuto dall'integrazione automatica di: WordNet, Wikipedia, OmegaWiki, Wiktionary, Wikidata, Wikiquote, VerbNet, Microsoft Terminology, GeoNames, ImageNet, FrameNet, WN-Map e Open Multilingual WordNet.

Essendo BabelNet un'estensione di WordNet e altri database, abbiamo un numero più elevato di synset per la lingua tedesca.

In questo studio, dunque, verranno usati i babel synset delle parole sia in lingua inglese sia in lingua tedesca.

2.4 spaCy

spaCy è una libreria open source gratuita per l'elaborazione avanzata del linguaggio naturale (NLP) in Python. Mentre alcune delle funzionalità di spaCy operano in modo indipendente, altre richiedono il caricamento di modelli statistici, che consentono a spaCy di prevedere le annotazioni linguistiche.

spaCy fornisce una varietà di annotazioni linguistiche per dare un'idea della struttura grammaticale di un testo. Ciò include i tipi di parola, come le componenti del discorso e il modo in cui le parole sono correlate tra loro. Durante l'elaborazione, spaCy prima tokenizza il testo, ovvero lo segmenta in parole, punteggiatura e così via. Questo viene fatto applicando regole specifiche per ogni lingua. Ogni documento è costituito da singoli token su cui possiamo iterare.

Dopo la tokenizzazione, spaCy può analizzare e taggare un determinato documento. È qui che entra in gioco il modello statistico, che consente a spaCy di fare una previsione di quale tag o etichetta più probabilmente si applica in questo contesto. Un modello viene prodotto mostrando a un sistema un numero sufficiente di esempi per poter compiere previsioni generalizzate in tutta la lingua.

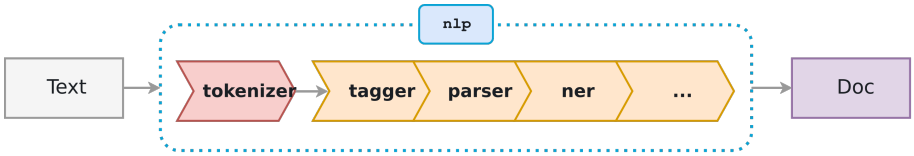


Figura 2.1. spaCy pipeline

spaCy può riconoscere vari tipi di entità nominate in un documento, chiedendo una previsione al modello. Poiché i modelli sono statistici e dipendono fortemente dagli esempi su cui sono stati addestrati, non sempre funziona perfettamente.

spaCy è in grado di confrontare due oggetti e prevedere di quanto siano simili.

Quando viene eseguito *nlp* su un testo, spaCy prima tokenizza il testo per produrre un oggetto *Doc*. Il documento viene quindi elaborato in diversi passaggi, denominati anche "pipeline di elaborazione". La pipeline utilizzata dai modelli predefiniti è costituita da un tagger, un parser e un riconoscimento di entità. Ogni componente della pipeline restituisce il documento elaborato, che viene quindi passato al componente successivo (*Figura 2.1*)

Sono queste funzionalità sulle annotazioni linguistiche che verranno utilizzate per lavorare sui documenti sia inglesi sia tedeschi.

2.5 spacy-babelnet

spacy-babelnet è una componente della pipeline spaCy che annota i token con i corrispondenti BabelNet Synset e con i lemmi. I synset vengono cercati solo nella lingua specificata, ma altre lingue possono essere, comunque, recuperate tramite BabelNet. Se il token ha un'annotazione POS, i synset vengono cercati solo con quel POS.

È la pipeline principale che viene usata in questa ricerca, verranno caricati i modelli grandi di spaCy sia per il tedesco sia per l'inglese. Le differenze di modello

sono per lo più statistiche. In generale, ci aspettiamo che i modelli più grandi siano complessivamente "migliori" e più precisi.

Capitolo 3

Criteri di valutazione

Possiamo definire i criteri di valutazione come le metriche utilizzate per calcolare l'efficienza degli algoritmi proposti. Verranno calcolati sia per l'algoritmo di allineamento sia in quello di trasferimento delle informazioni.

In entrambi viene definita un'ipotesi e vengono calcolati singolarmente i veri positivi (VP), casi che seguono l'ipotesi in modo corretto, i falsi positivi (FP), che seguono l'ipotesi in modo sbagliato, i veri negativi (VN), che non seguono l'ipotesi ma è giusto che non lo facciano e i falsi negativi (FN), che non seguono l'ipotesi ma dovrebbero farlo.

3.1 Precision

La precision è l'accuratezza con cui il sistema di classificazione prevede le classi positive, i veri e falsi positivi. È definito come il rapporto tra i veri positivi e la somma dei veri positivi e falsi positivi, ovvero il numero degli elementi classificati come riconosciuti. Quindi è la percentuale di riconoscimenti fatti esattamente:

$$\frac{VP}{VP + FP}$$

La sua definizione permette di ottenere una precisione del 100% avendo solamente una previsione positiva corretta oppure nessuna previsione falsa positiva.

Per questo motivo non viene mai considerata da sola, ma è sempre associata ad un'altra metrica, la recall.

3.2 Recall

La recall indica il rapporto tra i veri positivi e la somma dei veri positivi e falsi negativi, ovvero il numero di casi che si sarebbero dovuti riconoscere:

$$\frac{VP}{VP + FN}$$

Non vengono considerati solamente i risultati positivi ma anche quelli che hanno avuto un esito negativo nonostante ci si aspettasse il contrario.

3.3 F1

F1 score è una media armonica, cioè considera il reciproco della media aritmetica dei reciproci. Utilizza la precision e la recall e serve per misurare l'accuratezza di un test.

$$\frac{2}{\frac{1}{p} + \frac{1}{r}} = \frac{2pr}{p + r}$$

La media armonica attribuisce un peso maggiore ai valori piccoli. Questo permette di ottenere un alto F1 solo quando precisione e recupero sono entrambi alti.

Capitolo 4

Allineamento delle traduzioni Tedesco-Inglese

Prima di entrare nel merito della ricerca svolta bisogna spiegare cosa si intende con “allineamento” delle traduzioni. La scelta di questo termine non è casuale in quanto riesce ad esprimere al meglio ciò che si vuole rappresentare.

4.1 Definizione

Per capire cosa si ha intenzione di esprimere con il termine "allineamento" conviene far uso di un esempio:

Consideriamo la seguente frase inglese:

"It smells very delicious in the kitchen."

E le seguenti due frasi tedesche:

1. *"Es riecht sehr köstlich in der küche."*
2. *"Sie haben sich köstlich amüsiert."*

Allineare una frase in una lingua con una in un'altra lingua, significa semplicemente trovarne la giusta traduzione. In questo modo possiamo dire che la frase inglese nell'esempio è allineata con la prima frase tedesca, ma non con la seconda.

4.2 L'idea

L'idea alla base di questa ricerca è rendere il lavoro di allineamento delle traduzioni "automatizzato", cioè trasformare i testi così da renderli adatti ad essere trattati da un calcolatore, da usare poi per allineare le frasi.

Spieghiamo meglio ciò che intendiamo con un piccolo esempio.

Consideriamo il caso in cui una persona conosca l'inglese ma non il tedesco. Ha a disposizione i testi che gli interessano in entrambe le lingue. Sceglie una porzione del testo inglese di cui vuole conoscerne la parte corrispondente nel testo tedesco.

Osserviamo questa porzione di testo in inglese:

[...] Feeling a little ashamed of myself, I looked away from him, and from then on I avoided looking curiously at anything, even the tablecloth. ***Soon after that I said goodbye to my talkative friend. I noticed as I left that he immediately moved to the table where his military hero was sitting, probably to give him an account of me as eagerly as he had talked to me about Hofmiller.*** *That was all. [...]*

È di interesse capire la traduzione delle frasi evidenziate in grassetto. Di conseguenza, quello che succederà è che dal testo tedesco verrà selezionata la porzione corrispondente, come si può notare nel testo seguente:

[...] Gleichzeitig rückte jener Herr mit einer unverkennbar unfreundlichen Bewegung den Sessel zur Seite und schob uns energisch den Rücken zu. Etwas beschämt nahm ich meinen Blick zurück und vermied von nun an, auch nur die Decke jenes Tischs neugierig anzustreifen. ***Bald darauf verabschiedete ich mich von meinem braven Schwätzer, beim Hinausgehen jedoch schon bemerkend, daß er sich sofort zu seinem Helden hinübertransferierte, wahrscheinlich um einen ebenso eifrigen Bericht über mich zu erstatten wie zu mir über jenen.*** *[...]*

La persona invece di provare a trovare la traduzione "a mano", indubbiamente un lavoro troppo lungo specialmente se non si conosce una della due lingue, può dare i due testi in pasto all'algoritmo, che restituisce le corrispondenze delle frasi del testo inglese con quelle del testo tedesco.

Pertanto, come già detto, questo rende possibile, anche a chi non conosce entrambe le lingue, di recuperare le parti di testo di cui ha bisogno.

Dunque lo scopo principale è allineare le frasi di un romanzo inglese con le frasi del rispettivo romanzo tedesco e viceversa.

Occorre essere in possesso dei testi in entrambe le lingue per poterci lavorare, per questo motivo non deve essere considerato come un traduttore di documenti.

4.3 L'algoritmo

Prima di definire l'algoritmo finale, sono emerse diverse soluzioni, varianti in base ad alcuni aspetti analizzati successivamente.

L'idea generale non cambia, quindi prima di introdurre i diversi tentativi, spieghiamo i passi che sono stati fatti per produrre l'algoritmo conclusivo di questa ricerca.

Consideriamo i due testi: il primo passo da fare è darli in pasto alla pipeline di spacy-babelnet, per ottenere un *Doc*, il formato finale con cui è possibile lavorare un testo su un calcolatore.

Disponendo dei testi in questo nuovo formato, quello che ci interessa recuperare da essi sono i synset, il significato delle parole. Quello che facciamo è prendere tutti i synset delle parole di cui è composta ogni frase di ciascun testo. Ovviamente per ciascuna lingua vengono utilizzati i corrispondenti modelli di spaCy (sezione 2.4).

Sono i synset la parte fondamentale dell'algoritmo, essi ci aiuteranno a capire se una frase è effettivamente la traduzione dell'altra. È basandosi su questi che

abbiamo definito sei metodi differenti di valutazione di similarità. Questi metodi fanno uso di un valore di similarità che verrà usato come soglia, per cui, se il valore di similarità tra due frasi è uguale o supera questa soglia le frasi sono ritenute una la traduzione dell'altra, se invece non dovesse superarla allora non lo sono.

In tutti e sei i metodi è stata considerata una coppia di frasi, di cui una inglese e una tedesca. Sono di interesse i synset di ciascuna parola nella frase, infatti come numeratore del calcolo del valore di similarità è stato considerato l'insieme dei significati in comune tra queste due frasi. L'idea è che se nelle due frasi ci sono delle parole che hanno dei synset in comune è possibile che quei synset siano il significato corretto della parola, disambiguando le parole che in entrambe le lingue hanno più di un significato. Come denominatore, invece, ogni metodo ha il proprio che lo contraddistingue dagli altri:

1. *min_length*: conta quanti synset possiede ciascuna delle due frasi e prende il conteggio minore.

2. *max_length*: conta quanti synset possiede ciascuna delle due frasi e prende il conteggio maggiore.

3. *union*: conta quanti synset possiede ciascuna delle due frasi e li somma.

4. *max_words*: conta quante sono le parole che hanno almeno un synset per ciascuna delle due frasi e prende il conteggio maggiore.

5. *min_words*: conta quante sono le parole che hanno almeno un synset per ciascuna delle due frasi e prende il conteggio minore.

6. *sum_words*: conta quante sono le parole che hanno almeno un synset per ciascuna delle due frasi e li somma.

Il valore di similarità per ciascuna frase, quindi, viene ricavato dividendo il numeratore con il denominatore, in base al metodo che ci interessa.

Quindi una coppia di frasi avrà valori di similarità differente per metodo differente, per esempio:

le due frasi *"The ability to help others is a very important character trait."* e *"Die Fähigkeit, anderen zu helfen, ist ein sehr wichtiges Charaktermerkmal."* hanno come valore di similarità: 0.36 per *min_length*, 0.10 per *max_length*, 0.09 per *union*, 1.0 per *max_words*, 1.33 per *min_words* e 0.57 per *sum_words*.

Per ciascuno di questi metodi abbiamo calcolato la "soglia". Abbiamo usato 34 frasi in inglese e le 34 traduzioni tedesche. Di queste frasi, 15 sono state prese da giornali e articoli, mentre le altre sono state prese da romanzi. L'idea era quella di calcolare il valore di similarità tra le frasi con traduzione corrette e tra le stesse frasi con una traduzione presa a caso. Come mostrato in seguito:

è stato calcolato il valore di similarità, per esempio, tra la frase inglese *"They brightly decorated the wooden house for the holiday."* e le seguenti due frasi tedesche, la traduzione tedesca esatta *"Sie schmückten das Holzhaus hell für den Urlaub."* e una traduzione presa a caso *"Luke erklärte, warum Rosemary zu spät zum Unterricht kam."*. L'obiettivo è di non dare un peso maggiore alle frasi che sono esattamente la traduzione, ma proporzionarle al caso in cui non ci dovesse essere la traduzione.

Applicando questo concetto alle 34 frasi, per ciascuna di esse abbiamo ottenuto un valore per quando la traduzione era quella esatta e un valore nel caso in cui fosse quella sbagliata. Abbiamo fatto la media considerando tutte le frasi per entrambi i casi, ottenendo quindi due valori, uno che rappresenta la soglia per le frasi tradotte correttamente e uno per quelle tradotte in modo errato. Dunque, per ogni metodo, abbiamo due valori:

- per le frasi con traduzione corretta:

  min_length: 0.299 *union*: 0.062

  max_length: 0.076 *max_words*: 1.129

min_words: 1.549 *sum_words*: 0.639

- per le frasi con traduzione casuale:

min_length: 0.007 *max_words*: 0.016

max_length: 0.001 *min_words*: 0.049

union: 0.001 *sum_words*: 0.012

La "soglia" è la media tra questi due valori per ciascun metodo, cioè:

min_length: 0.153 *max_words*: 0.573

max_length: 0.039 *min_words*: 0.799

union: 0.031 *sum_words*: 0.324

Come già accennato precedentemente, questi valori sono stati utilizzati per definire la corrispondenza tra le frasi tedesche e inglesi.

Conoscendo pertanto il criterio con cui vengono valutate le frasi, possiamo introdurre i diversi tentativi di elaborazione dell'algoritmo.

4.3.1 Primo tentativo

Per calcolare la similarità tra due frasi abbiamo considerato un controllo uno ad uno. Cioè ogni frase inglese viene confrontata con una sola frase tedesca, in ordine. Quindi la prima frase inglese viene confrontata con la prima tedesca ricavando un valore di similarità, la seconda inglese con la seconda tedesca e così via.

È stato preso in considerazione questo modo di lavorare con le frasi, poiché è solito avere una traduzione in ordine dall'inglese al tedesco e viceversa. Anche se risulta essere in ordine, sono stati tanti i casi in cui una frase tedesca nella sua traduzione inglese veniva spezzata in più frasi, e viceversa, e ci possono essere casi in

cui manca proprio la traduzione. Di conseguenza, considerando una frase per volta, la frase tedesca non avrebbe avuto la traduzione corretta. Come mostrato sotto:

le due frasi inglesi:

Brother Otho hastened from his library and I from the herbarium on the inner gallery. Lampusa, too, forsook her fireside to watch the child with proud and tender look.

corrispondono ad un'unica frase tedesca:

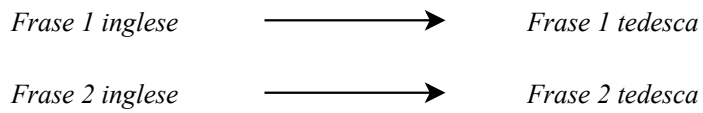
Bruder Otho eilte aus seiner Bibliothek und ich aus dem Herbarium auf den inneren Altan, und auch Lampusa trat vom Herd hinzu und lauschte dem Kinde mit stolzem, zärtlichem Gesicht.

Dunque, si sono verificati molti più casi in cui la traduzione non era corretta, dal momento che a partire dalla prima frase risultata non traduzione, di conseguenza tutte le altre non lo sarebbero state.

4.3.2 Secondo tentativo

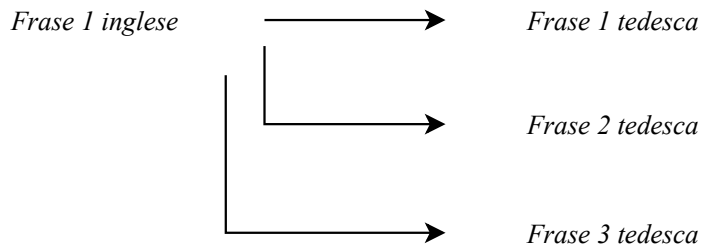
Nel secondo tentativo abbiamo considerato un controllo una a molti. Al posto di controllare due frasi e ricavarne il valore di similarità, abbiamo distinto questo controllo in tre casi. Vediamo con uno schema cosa si intende:

- con k = 1, cioè consideriamo solo una frase alla volta da cui prendere il valore:



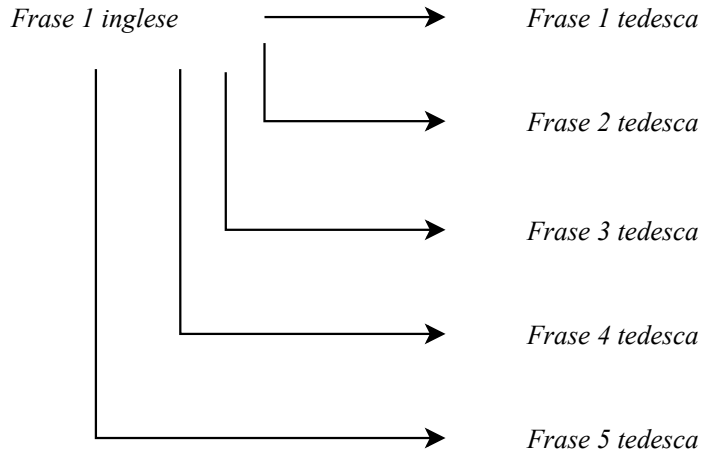
corrisponde al caso analizzato nel primo tentativo.

- con k = 3:



una singola frase inglese alla volta viene confrontata con 3 frasi tedesche successive, però solo una di queste verrà scelta come traduzione, quella con il valore di similarità maggiore.

- con k = 5:



come con k = 3, solo che le singole frasi vengono confrontate con 5 frasi nell'altra lingua.

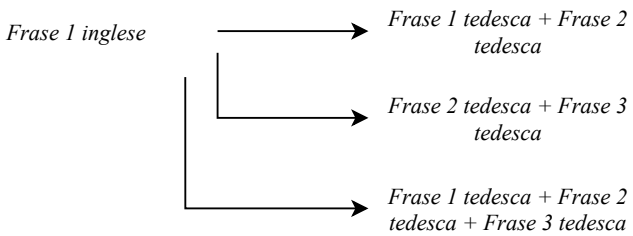
In primo luogo, abbiamo confrontato una frase inglese considerando una sola frase tedesca per tutte le frasi, poi con tre e cinque frasi, prendendo negli ultimi due casi, la frase che avesse maggiore valore di similarità. Il confronto quindi viene fatto sempre uno ad uno, senza però considerare solo la corrispondente frase nell'altra lingua, ma anche le successive. Abbiamo deciso di tenere il caso k = 1, nonostante al primo tentativo ci abbia dato problemi, poiché, come già accennato, molti testi sono tradotti in ordine.

Nonostante ciò, in questo tentativo non è stato considerato il caso in cui delle frasi venissero spezzate in più frasi nell'altra lingua.

4.3.3 Tentativo Finale

Il tentativo finale segue lo stesso approccio del secondo tentativo con k = 1, 3, 5, ma sono state considerate anche le frasi spezzate. In sostanza, quando si confrontano le frasi, si ricava il valore di similarità attraverso il calcolo visto precedentemente, per ciascuno dei 6 metodi: se questo valore è maggiore o uguale alla soglia ricavata per il metodo corrispondente, allora è ritenuta traduzione di quella frase, se invece non dovesse superarla, abbiamo considerato che potesse essere stata spezzata. Di conseguenza al posto di confrontare le singole frasi, per la lingua in cui abbiamo considerato k frasi, proviamo a fondere quelle che sono una la successiva dell'altra, cioè:

con k = 3, per esempio, avremmo questo controllo:



Per ricalcolare il valore di similarità si sommano i synset delle frasi unite, trattandole come se fossero una frase unica. Se, facendo di nuovo il controllo, una di queste unioni supera la soglia, quella viene considerata traduzione della frase, altrimenti indichiamo la frase come se non avesse una traduzione, in quanto né confrontando le frasi singolarmente né facendone l'unione, il valore di similarità supera la soglia.

Non trovandoci esclusivamente nel caso 1 ad 1, fare il confronto tra una frase inglese e k tedesche è differente da confrontare una frase tedesca e k inglesi.

Per questo motivo, abbiamo deciso di fare un confronto per entrambe le direzioni, cioè oltre il verso visto negli esempi precedenti consideriamo anche che la frase singola è quella tedesca e le k frasi sono quelle inglesi, e fondiamo i risultati in modo da non dare priorità a nessuna delle due direzioni.

4.4 Analisi del tentativo finale

Sono stati analizzati porzioni di 6 testi: *Heart of Darkness* di Joseph Conrad, *The masque of the Red Death* di Edgar Allan Poe, *Treasure Island* di Robert Louis Stevenson, *Siddhartha* di Hermann Hesse, *Ungeduld des Herzens* di Stefan Zweig e *Auf den Marmorklippen* di Ernst Jünger.

Definiamo l'ipotesi per la quale sono state valutate le analisi fatti:

"Data una frase in una delle due lingue, questa ha una corretta corrispondenza nell'altra lingua."

Dunque abbiamo:

- un *vero positivo*, se la frase ha una corrispondenza e questa è corretta.
- un *falso positivo*, se la frase ha una corrispondenza, ma è quella sbagliata.
- un *vero negativo*, se la frase non ha corrispondenza ed è giusto che non la abbia.

- un *falso negativo*, se la frase ha una corrispondenza, ma non viene trovata.

Sono state analizzate un totale di 746 frasi considerando entrambi le direzioni, di cui:

per l'inglese: 42 sono di *Heart of Darkness*, 84 di *The masque of the Red Death*, 35 di *Treasure Island*, 52 di *Siddhartha*, 63 di *Ungeduld des Herzens*, 103 di *Auf den Marmorklippen*.

per il tedesco: 42 sono di *Heart of Darkness*, 80 di *The masque of the Red Death*, 38 di*Treasure Island*, 53 di *Siddhartha*, 51 di *Ungeduld des Herzens*, 103 di *Auf den Marmorklippen*.

Per analizzare i testi è stata usata una stringa che contiene l'allineamento "perfetto", cioè in cui sono presenti le corrispondenze esatte tra le frasi nelle due lingue. È stata creata per ciascun testo e ciascuna direzione.

Una stringa del genere viene generata anche dall'algoritmo, ma contiene le corrispondenze trovate applicandolo sui testi.

Per definire, quindi, la correttezza delle corrispondenze, viene confrontata la stringa uscita dall'algoritmo, per ciascun metodo e direzione, con quella "perfetta".

Abbiamo voluto distinguere i risultati in base alla k, alla direzione e ai metodi, come possiamo notare nei rispettivi grafici in *Figura 4.1*, *Figura 4.2* e *Figura 4.3*.

Il risultato finale comprende, comunque, tutti questi aspetti.

Secondo i dati utilizzati in questa ricerca:

- La k più accurata è k = 1, quindi il caso in cui consideriamo un controllo uno ad uno.

- La direzione che ci dà dei risultati migliori è dal tedesco all'inglese, in quanto è più probabile trovare delle frasi tedesche più lunghe che in inglese vengono suddivise in più frasi. Nei risultati finali i dati di queste due direzioni vengono raggruppate.

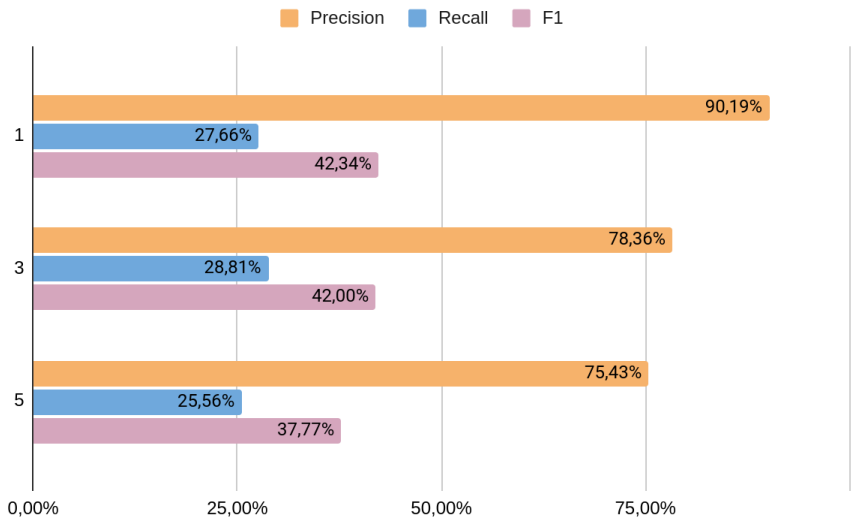


Figura 4.1. Analisi solo sulle k

- Il metodo migliore è quello che considera l'unione di tutti i synset delle frasi, *union*.

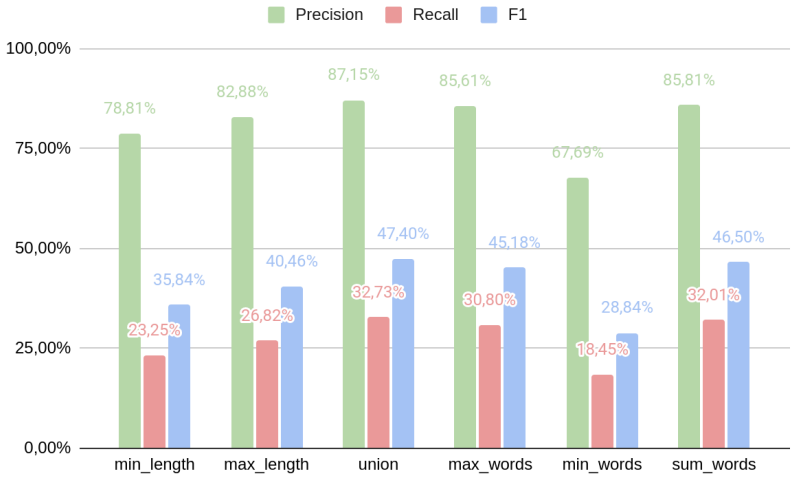


Figura 4.3. Analisi solo sui metodi

4.5 Risultati

Considerando tutti questi aspetti nell'insieme, i due metodi migliori, cioè quelli che sono risultati più adatti al nostro caso, sono *sum_words* e *union*, come mostrato

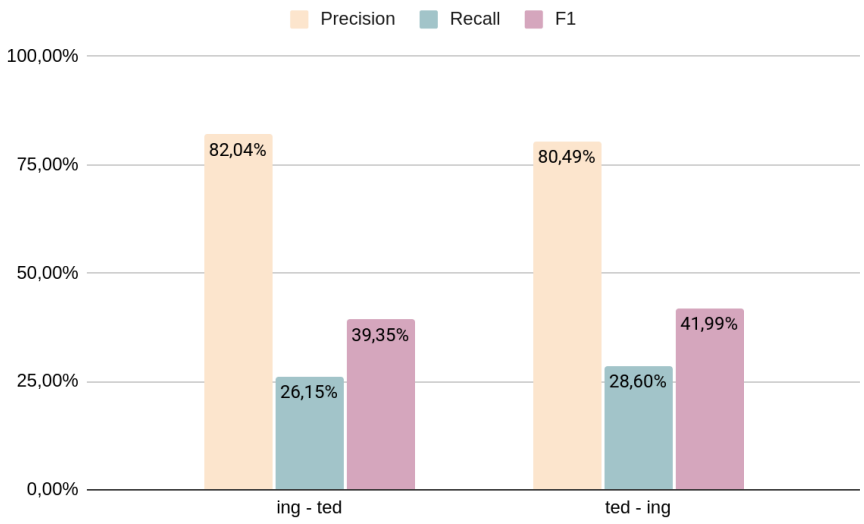


Figura 4.2. Analisi solo sulla direzione

nella tabella in *Figura 4.4*, in cui possiamo vedere le percentuali di tutti i metodi con tutte le k.

	Metodo	k	F1	Precision	Recall
1	union	5	49,85%	84,83%	35,29%
2	union	3	49,85%	85,71%	35,14%
3	sum_words	5	48,73%	85,05%	34,14%
4	sum_words	3	48,26%	83,10%	34,01%
5	max_words	5	47,17%	83,88%	32,81%
6	max_words	3	46,54%	83,03%	32,33%
7	max_length	3	43,34%	81,03%	29,58%
8	min_length	1	42,51%	90,91%	27,74%
9	union	1	42,51%	90,91%	27,74%
10	min_words	1	42,51%	89,29%	27,89%
11	sum_words	1	42,51%	89,29%	27,89%
12	max_length	1	42,17%	90,83%	27,46%
13	max_words	1	41,84%	89,91%	27,26%
14	min_length	3	36,38%	74,66%	24,05%
15	max_length	5	35,88%	76,78%	23,41%
16	min_length	5	28,64%	70,86%	17,95%
17	min_words	3	27,64%	62,63%	17,73%
18	min_words	5	16,36%	51,16%	9,73%

Figura 4.4

Analizzando la tabella si può notare che la recall è molto bassa, ciò significa che delle frasi che avrebbero dovuto avere la traduzione sono state indicate come se non

l'avessero.

È stato osservato che la quantità di falsi negativi è minore con k = 3 e in direzione tedesco-inglese. Ciò significa che dovremmo usare un range di frasi non troppo grande, superiore comunque alla singola frase. La direzione ci indica come è più probabile che le frasi vengano spezzate in inglese. Infatti se nella direzione inglese-tedesco si considerasse anche un confronto molti a molti, molto probabilmente il numero di falsi negativi sarebbe minore.

4.6 Funzionalità supplementari

4.6.1 Sentence splitter

È sorto un problema quando i testi venivano passati dalla pipeline di spaCy. Una delle funzionalità di quest'ultima è quella di separare il testo in frasi.

Nonostante ciò, non tutte le frasi sono state spezzate in modo corretto: sono stati rilevati dei casi in cui una frase veniva divisa dove si trovava una virgola oppure un simbolo di escape, come il ritorno a capo. Le frasi si possono presentare in quattro maniere differenti:

1. è completa, dunque è stata spezzata bene e non ci sono problemi.

   "The surgeon had been sitting with his face turned towards the fire: giving the palms of his hands a warm and a rub alternately."

2. è l'inizio di una frase, in quel caso viene aggiunta la frase successiva.

   "The surgeon had been sitting with his face turned towards the fire:"

3. è intermedia, cioè non è né l'inizio di una frase né la fine, in quel caso si univa sia alla precedente sia alla successiva.

   "his face turned towards the fire:"

4. è la fine della frase, in quel caso viene aggiunta alla frase precedente.

"giving the palms of his hands a warm and a rub alternately."

Per questo motivo è stata utilizzata una funzione che ricontrolla le frasi e che le sistema come suggerito sopra. Un motivo per cui lo spezzamento di spaCy non è stato fatto correttamente può essere legato al formato del testo dato in input. Non sempre tutti i caratteri sono supportati quando si passa da un formato all'altro. Quindi un passo fondamentale per evitare di dover ricontrollare le frasi e aggiungere questa funzionalità è quello di fare una pulizia del testo.

4.6.2 Discorso diretto

Dobbiamo trattare il discorso diretto a parte, in quanto risulta essere più difficile lavorarci, infatti sono stati rilevati dei problemi durante l'esecuzione dell'algoritmo. Nel nostro caso sono stati scelti porzioni di testi senza la presenza di dialoghi. Ciononostante è stato definita una funzionalità che permette di trattarli come delle singole frasi, per esempio:

"Young master," he said consolingly, "don't be sad. Most everyone is a little homesick at first, for his father, his mother, his brothers and sisters. But you'll see: life isn't bad here either, not bad at all."

È interpretata come unica frase anche se è presente un punto di interruzione, in quanto è parte del discorso.

L'algoritmo proposto funzionerebbe anche col discorso diretto, ma esclusivamente nel caso in cui non ci siano più "virgolette" del solito, riprendiamo la frase di prima adattandola al nostro caso:

"Young master," he said consolingly, "don't be sad.
"Most everyone is a little homesick at first, for his father, his mother, his brothers and sisters. But you'll see: life isn't bad here either, not bad at all."

Esistono situazioni in cui il discorso continua nella riga successiva e per far capire il suo continuo si è soliti aggiungere un'altra virgoletta ad inizio frase, rendendo sproporzionato il conteggio delle virgolette.

Questo caso non è stato gestito nell'attuale algoritmo, motivo per cui, come detto precedentemente, è stato deciso di considerare porzioni di testo senza la presenza di parti discorsive, ciò non esclude il fatto che non sia possibile farlo.

Capitolo 5

Pronominal Coreference

Dato un romanzo originale tedesco e la sua traduzione inglese, quello che vogliamo aggiungere al testo tedesco è l'informazione di "pronominal coreference". Ma cos'è?

5.1 Definizione

La coreference è una relazione che lega due o più espressioni che si riferiscono alla stessa entità.

Prendiamo un esempio per capire meglio cosa si intende:

"Bob said that he would go"

Il sostantivo *Bob* e il pronome *he* si riferiscono alla stessa persona.

Dobbiamo introdurre altri due termini importanti nel contesto della coreference: l'anafora e la catafora. Abbiamo una relazione di anafora quando un termine, "l'anafora", si riferisce ad un altro termine, "l'antecedente", e l'interpretazione del primo dipende da quella del secondo.

Billy is in the kitchen. He is cooking.

L'anafora è "he" e l'antecedente è "Billy".

Viene definita una relazione di catafora quando un termine, "la catafora", si riferisce ad un altro termine e l'interpretazione del secondo dipende da quella del primo. Può essere perciò definita l'opposto dell'anafora pur essendo meno diffusa .

I aet it. The cake was delicious.

"The cake" e "it" indicano la stessa cosa, ma il pronome viene definito prima.

5.2 L'idea

L'idea è quella di trasferire delle informazioni da un testo ad un altro semplicemente stabilendo a priori che uno è la traduzione dell'altro. Se ad una determinata frase di una lingua corrisponde una frase in un'altra lingua, possiamo prendere le informazioni che sono contenute in una e trasmetterle all'altra.

Il lavoro che viene svolto in questa parte della ricerca, quindi, non è di interpretazione del testo, ma è un semplice passaggio di informazioni.

Affinché questo trasferimento sia efficace, almeno uno dei due documenti ha bisogno di contenere le informazioni d'interesse.

È in questo contesto che entra in gioco BookNLP. Come spiegato nella sezione 2.1 del Capitolo 2, BookNLP si occupa principalmente di analizzare i personaggi in un testo inglese. Pertanto, è il romanzo inglese ad essere dato in pasto a BookNLP. Le informazioni del testo vengono interpretate e analizzate seguendo la pipeline di BookNLP, tra gli output viene generato un testo contenente le coreference. Queste informazioni vengono trasferite nel romanzo tedesco.

Le coreference trasmesse come quelle ricavate da BookNLP sono esclusivamente quelle in cui i pronomi personali e possessivi sono riferiti a personaggi (nomi propri) di terza persona singolare maschile e femminile (*he, she, him, his, her*).

Poiché le informazioni non vengono analizzate nella frase tedesca, ma solo trasferite, è stato deciso di considerare solo le traduzioni corrispondenti ai pronomi

e non di interpretarle nel contesto. Quindi non tutte le pronominal coreference vengono trasferite, in quanto possono esserci dei casi in cui un pronome che in una lingua è un pronome personale associato al soggetto, nell'altra può essere interpretato come un pronome personale associato all'oggetto e viceversa.

Prendiamo per esempio questa frase tedesca:

Mehr als sie alle aber liebte ihn Govinda, sein Freund, der Brahmanensohn.

Il pronome personale "ihn" viene comunemente tradotto nell'inglese "him". Ma il testo in inglese è stato interpretato diversamente:

But more than all the others he was loved by Govinda, his friend, the son of a Brahman.

Al posto di "him" abbiamo "he", quindi, come detto, non analizzando la frase ma trasferendo le informazioni, questa coreference verrebbe persa.

In questo capitolo vengono analizzati i testi che hanno un "allineamento perfetto", cioè in cui sappiamo a quale frase inglese corrisponde quella tedesca.

In questo modo intendiamo anticipare quello che verrà trattato nel capitolo successivo (Capitolo 6), mostrando cosa succede nel caso in cui tutte le frasi dovessero risultare allineate con la traduzione. Nel prossimo capitolo verranno applicate queste stesse funzioni, con qualche variante, solo alle frasi che sono risultate "positive" dal lavoro svolto nel Capitolo 4.

5.3 L'algoritmo

Analizziamo tutti i passi che sono stati compiuti per trasferire le informazioni.

Per prima cosa, abbiamo bisogno di usare BookNLP per ricavare il testo con le coreference. Prendiamo il romanzo originale tedesco tradotto in inglese e lo diamo in input a BookNLP. Ci serviamo del file html generato in output.

Questo file è diviso in due parti, abbiamo la sezione "Characters" che contiene un elenco dei personaggi ricorsi nel testo, in cui sulla stessa riga ritroviamo i diversi modi in cui un personaggio viene chiamato nel testo, e la sezione "Text" che contiene il testo vero e proprio con le coreference associate ai pronomi.

Questo file viene trasformato in due file di testo distinti, uno conterrà i personaggi, mentre l'altro il testo.

Prima di poter eseguire l'algoritmo, dobbiamo sapere quali sono le corrette corrispondenze tra le frasi tedesche e quelle inglesi.

Trattandosi del caso "perfetto" abbiamo utilizzato la stringa creata per analizzare la correttezza dell'algoritmo del Capitolo 4. Questa stringa indica a quali frasi tedesche corrispondono le frasi inglesi.

L'algoritmo di trasferimento delle informazioni prende in input i due testi, quello inglese con le coreference e quello tedesco. Considera le coppie di frasi tedesche e inglesi che sono risultate essere la traduzione reciproca e controlla quanti e quali sono i pronomi con coreference nella frase inglese e le associa ai corrispondenti pronomi della frase tedesca.

Abbiamo considerato che le due frasi abbiano i pronomi in ordine, cioè se nella frase inglese troviamo, in ordine, *he*, *his* e *she*, ci aspettiamo di trovare *er*, *seines* e *sie* nella frase tedesca. Ciò si è verificato in tutte le frasi analizzate in questa ricerca; tuttavia questo potrebbe non valere sempre.

5.4 Analisi

Sono stati analizzati porzioni di tre romanzi originali tedeschi, *Siddhartha* e *Narziß und Goldmund* di Hermann Hesse, e *Auf den Marmorklippen* di Ernst Jünger.

Nonostante non siano stati considerati dei testi interi, è meglio darli comunque in input a BookNLP, poiché, come detto precedentemente, una coreference potrebbe

essere identificata sia in una relazione di anafora sia in una di catafora.

Definiamo l'ipotesi per la quale sono stati valutati le analisi fatte:

"Ad ogni pronome personale o possessivo di terza persona singolare maschile e femminile presente nel testo tedesco è associato il personaggio corretto."

Dunque abbiamo:

- un *vero positivo*, se dato un pronome, gli è stato associato un personaggio ed è quello corretto.
- un *falso positivo*, se dato un pronome, gli è stato associato un personaggio, ma è quello sbagliato.
- un *vero negativo*, se dato un pronome non gli è stato associato nessun personaggio e non doveva esserci.
- un *falso negativo*, se dato un pronome non gli è stato associato nessun personaggio ma avrebbe dovuto averlo.

Sono state analizzate un totale di 181 frasi, di cui 52 di *Siddhartha*, 26 di *Narziß und Goldmund* e 103 di *Auf den Marmorklippen.*

Nelle 52 frasi sono presenti un totale di 82 pronomi, di questi sono risultati 29 veri positivi, 10 veri negativi, 41 falsi positivi e 2 falsi negativi.

Nelle 26 frasi sono presenti un totale di 29 pronomi, di questi sono risultati 10 veri positivi, 1 vero negativo, 1 falso positivo e 17 falsi negativi.

Nelle 103 frasi sono presenti un totale di 35 pronomi, di questi sono risultati 13 veri positivi, 0 veri negativi, 11 falsi positivi e 11 falsi negativi.

Nella *Figura 5.1* è rappresentato il grafico con le percentuali dei positivi e negativi per ciascun testo.

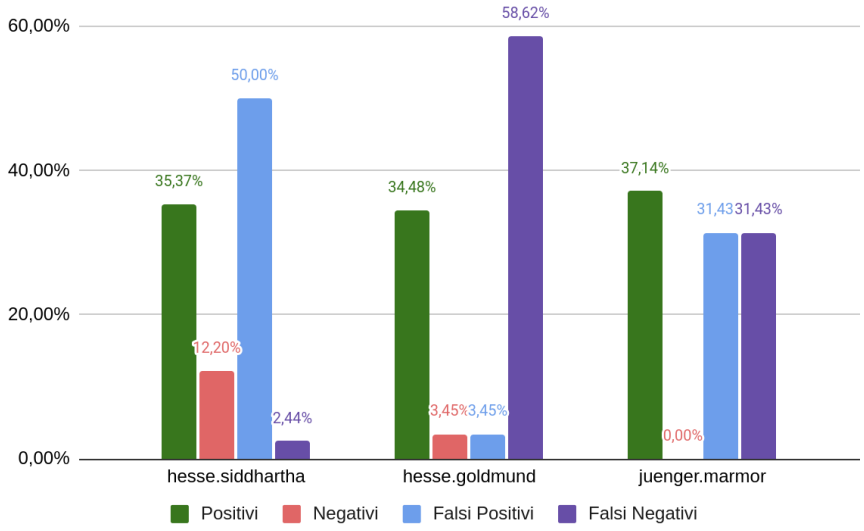


Figura 5.1

5.4.1 Risultati

Per calcolare la precision, la recall e l'F1, sono stati considerati i tre testi come se fossero un testo unico. Dunque, con un totale di 146 pronomi, abbiamo 52 veri positivi, 11 veri negativi, 53 falsi positivi e 30 falsi negativi, abbiamo ottenuto i seguenti risultati:

Precision	Recall	F1
49.52%	63.41%	55.61%

Essendo presenti molti falsi positivi, cioè casi in cui la coreference è presente ma il personaggio è quello sbagliato, un modo per migliorare l'algoritmo è analizzare questi casi e rendere l'algoritmo più selettivo.

Capitolo 6

Allineamento per trasferire informazioni di coreference

In questo capitolo verrà trattato un caso particolare di trasferimento delle informazioni visto nel capitolo precedente.

Non ci troviamo più a trattare con un "allineamento perfetto", ovvero quello in cui siamo certi che ad una determinata frase in una lingua corrisponda la traduzione nell'altra, ma con un nuovo caso, per cui verranno prese in considerazione tutte e sole le frasi che sono risultate essere delle traduzioni, cioè "positive", nei due casi migliori di allineamento visto nel Capitolo 4.

L'idea non è differente da quella con l'allineamento perfetto ma ricapitoliamo brevemente i passaggi.

Viene considerato il testo inglese con coreference ricavato da BookNLP e il testo corrispondente tedesco, dal primo vengono trasferite le informazioni al secondo. Questo passaggio viene eseguito considerando le frasi che risultano essere corrispondenti tra loro e confrontando i pronomi tra esse.

L'unica variazione è presente nell'algoritmo, in cui viene fatto un controllo che prende solo le frasi con corrispondenza positiva uscite dai metodi *union* e *sum_words*,

entrambi con $k = 5$.

6.1 Analisi

Sono stati analizzati tre dei sei testi considerati nel Capitolo 4, gli originali tedeschi, *Siddhartha* di Hermann Hesse, *Auf den Marmorklippen* di Ernst Jünger e *Ungeduld des Herzens* di Stefan Zweig.

Riproponendo la stessa ipotesi del Capitolo 5: *"Ad ogni pronome personale o possessivo di terza persona singolare maschile e femminile presente nel testo tedesco è associato il personaggio corretto."*

In cui abbiamo:

- un *vero positivo*, se dato un pronome, gli è stato associato un personaggio ed è quello corretto.
- un *falso positivo*, se dato un pronome, gli è stato associato un personaggio, ma è quello sbagliato.
- un *vero negativo*, se dato un pronome non gli è stato associato nessun personaggio e non doveva esserci.
- un *falso negativo*, se dato un pronome non gli è stato associato nessun personaggio ma avrebbe dovuto averlo.

Abbiamo studiato i due metodi separatamente, ottenendo valori differenti:

per *union*: su 76 frasi totali, sono stati individuati 38 pronomi, di questi 11 sono veri positivi, 4 sono veri negativi, 17 sono falsi positivi e 6 sono falsi negativi.

per *sum_words*: su 71 frasi totali, sono stati individuati 46 pronomi, di questi 14 sono veri positivi, 6 sono veri negativi, 16 sono falsi positivi e 10 sono falsi negativi.

Affinché fossero evidenti eventuali errori nel trasferimento, sono state confrontate le frasi con le rispettive frasi corrette e non con quelle che gli sono state associate dal metodo di allineamento.

Abbiamo potuto notare che, nel caso in cui la traduzione non fosse quella corretta, alcuni pronomi non sarebbero stati trasferiti: infatti, è molto improbabile che nella frase ci sia il pronome corretto, e se per caso ci fosse è difficile che vi sia associato il personaggio corretto.

Nonostante ciò abbiamo riscontrato dei casi in cui, nella traduzione corretta sono stati trovati dei falsi positivi, che invece con quella errata sono diventati dei veri positivi. Ovviamente questa eventualità non è frequente, nella maggior parte dei casi, quando la traduzione non risultava quella corretta, non sono state trasferite le coreference, perciò risultano essere dei falsi negativi.

Consideriamo un esempio che comprende entrambi i casi spiegati precedentemente. Analizziamo delle frasi dei testi considerati in questa parte della ricerca:

in seguito è riportata la frase inglese con coreference:

And in days to come, when Siddhartha would become a god, when he (Siddhartha) would join the glorious, then Govinda wanted to follow him (Govinda) as his (Govinda) friend, his (Govinda) companion, his (Govinda) servant, his (Govinda) spear-carrier, his (Govinda) shadow.

Sono presenti 7 pronomi con coreference, di cui il primo è corretto e gli altri 6 sono errati.

Consideriamo ora i due trasferimenti:

- nel caso di allineamento perfetto:

  Und wenn Siddhartha einstmals ein Gott würde, wenn er (Siddhartha) einstmals eingehen würde zu den Strahlenden, dann wollte Govinda ihm

(Govinda) folgen, als sein (Govinda) Freund, als sein (Govinda) Begleiter, als sein (Govinda) Diener, als sein (Govinda) Speerträger, sein (Govinda) Schatten.

- nel caso di allineamento dal metodo *sum_words*:

Und wenn Siddhartha einstmals ein Gott würde, wenn er (Siddhartha) einstmals eingehen würde zu den Strahlenden, dann wollte Govinda ihm folgen, als sein (Siddhartha) Freund, als sein (Siddhartha) Begleiter, als sein Diener, als sein Speerträger, sein Schatten.

Come possiamo notare, nel primo caso di trasferimento, vengono associati a tutti i pronomi le coreference, i 7 pronomi con coreference nell'inglese ai corrispondenti 7 pronomi in tedesco. Il trasferimento viene eseguito correttamente, ma sui 7, come per l'inglese, solo 1 è un vero positivo, mentre gli altri 6 sono dei falsi positivi.

Mentre nel secondo caso di trasferimento, solo 3 dei 7 pronomi hanno una coreference associata, ma tutti e tre sono corretti.

È stata soltanto una casualità, ma la frase sbagliata che gli è stata associata è risultata avere più coreference corrette di quella effettivamente giusta.

6.1.1 Risultati

I risultati sulla precision, la recall e l'F1 sono stati calcolati per ciascun metodo e sono rappresentati nel grafico in *Figura 6.1*

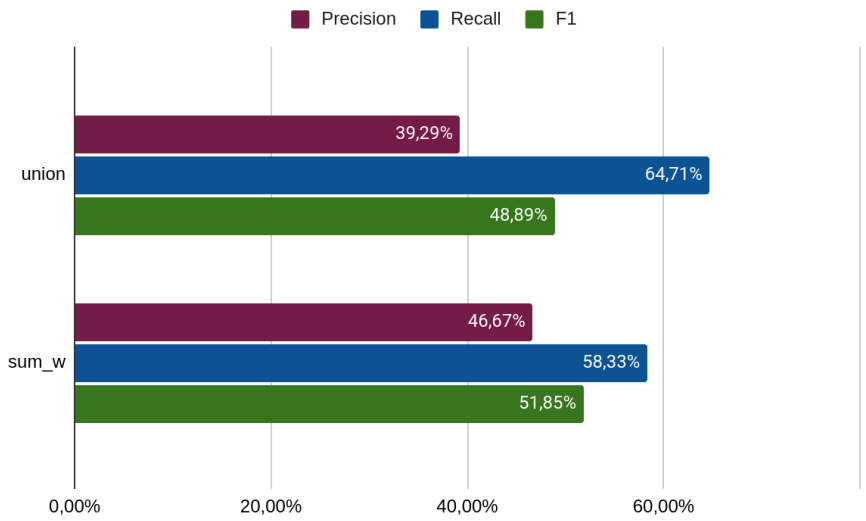


Figura 6.1

Dati i risultati, possiamo affermare che tra i due, *sum_words* risulta essere il metodo più adatto.

Il numero di frasi analizzate non è elevato, perciò sia con un numero di frasi diverso sia con romanzi differenti, il metodo risultato più efficiente in questo caso potrebbe non esserlo in altri.

Capitolo 7

Conclusione

Tutte le scelte pensate durante lo sviluppo di questa ricerca, sono state prese in quanto ritenute quelle più pertinenti al momento, nonostante ciò non si può escludere che alcuni degli approcci presi in considerazione non siano perfetti e che ci siano altri modi di approcciarsi a questi tipi di problemi.

Per quanto riguarda l'algoritmo di allineamento, questo può essere migliorato in molti punti, tra i quali è doveroso considerare la possibilità di includere porzioni di testo più grandi o interi paragrafi che trattano di un argomento simile, oppure considerare non solo un controllo uno a uno e uno a molti, ma anche molti a molti. Un altro aspetto è, per esempio, di non prendere in esame solo i synset per decidere se una frase è l'allineamento dell'altra. Avendo osservato che molte frasi che avrebbero dovuto avere una traduzione non hanno avuto corrispondenza, ciò che potrebbe essere cambiato è la soglia. Abbassandola è più probabile di trovare che due frasi sono la rispettiva traduzione.

Per quanto riguarda l'algoritmo di trasferimento, può essere migliorato cercando di interpretare minimamente le frasi per cercare di non sbagliare ad associare una coreference ad un pronome e capire quando è plurale o singolare, in quanto in tedesco in alcuni casi non c'è distinzione se non quando si va ad interpretare la frase, oppure

quando un pronome è stato considerato come riferimento ad oggetto o soggetto. In questo modo, si potrebbero evitare i falsi negativi.

Oppure, se si considera un paragrafo, vedere se è presente il soggetto, anche nelle frasi precedenti, e confrontarlo con quello indicato dalla frase inglese, in modo da evitare falsi positivi.

Bibliografia

[1] David Bamman, Ted Underwood and Noah Smith, "A Bayesian Mixed Effects Model of Literary Character," ACL 2014.

[2] Purva Huilgol, "Quick Introduction to Bag-of-Words (BoW) and TF-IDF for Creating Features from Text", 2020

[3] Şaziye Betül Özateş, Arzucan Özgür, Dragomir R. Radev, “Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization”

[4] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D.Manning, "Generating Typed Dependency Parses from Phrase Structure Parses"

[5] Princeton University "About WordNet." WordNet. Princeton University. 2010.

[6] Stanford University "Stanford CoreNLP: A Java suite of core NLP tools." Stanford CoreNLP.

[7] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations", pp. 55-60.

[8] Roberto Navigli and Simone Paolo Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network"

[9] spaCy, "spaCy 101: Everything you need to know", spaCy

[10] Honnibal, Matthew and Montani, Ines "Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing" 2017

[11] asterbini, "spacy-babelnet", GitHub 2020

[12] Andrea Provino, "Precision and Recall con F1 Score | Precisione e Recupero" 2019