HOMEWORK 4

ALESSIO MARINUCCI

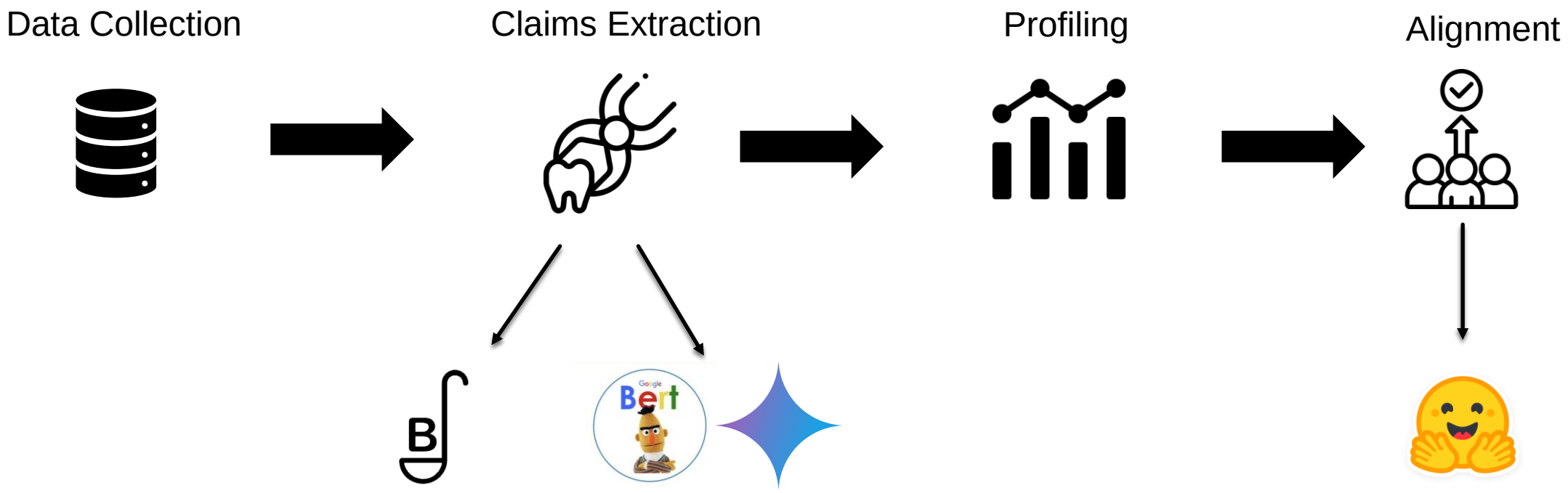
RICCARDO FELICI

https://github.com/DrRicky31/HW4_IID

OBIETTIVO DELLO STUDIO

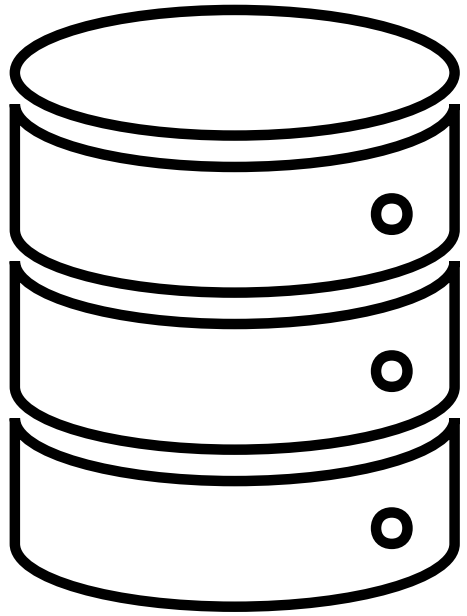
- Estrazione di informazioni strutturate da un insieme di file JSON contenenti tabelle presenti in articoli scientifici.
- Tali informazioni saranno estratte direttamente dalle tabelle HTML e arricchite con dati contestuali come didascalie, riferimenti e note a piè di pagina.
- Il processo di estrazione utilizza tecniche avanzate di elaborazione del linguaggio naturale per identificare dati significativi.
- Successivamente i dati strutturati vengono sottoposti a un'analisi statistica per calcolare distribuzioni, medie e altre statistiche descrittive per individuare trend o anomalie.
- Infine viene effettuato un allineamento terminologico per standardizzare e unificare termini utilizzati nelle tabelle.

PIPELINE


Data Collection
Claims Extraction
Profiling
Alignment
Bert

DATA COLLECTION

- Sono state collezionate 37 tabelle a partire da 13 papers scelti riguardo il topic di "*Machine Traslation*".
- E' stata fatta una classificazione di tabelle, per una gestione più efficiente, in base al tipo di informazioni contenute all'interno:
  - Tipo 1: tabelle relazionali, con l'utilizzo di LLM per distinguere metriche e specifiche contenute nelle colonne.
  - Tipo 2: tabelle che necessitano di una estrazione di informazioni aggiuntive contenute nella caption e nei paragrafi con riferimenti alla tabella.
  - Tipo 3: tabelle nidificate che si comportano come le tabelle di tipo 1.



DATA COLLECTION

Dataset	ε	Iteration Method	Test BLEU	Test BERTScore
WMT-16	∞	Random shuffling	36.19 (0.13)	0.95 (0.00)
WMT-16	1000	Random shuffling	20.86 (0.56)	0.92 (0.00)
WMT-16	1000	Poisson sampling	15.12 (0.08)	0.91 (0.00)
WMT-16	5	Random shuffling	19.24 (0.52)	0.92 (0.00)
WMT-16	5	Poisson sampling	7.23 (0.21)	0.89 (0.00)
WMT-16	1	Random shuffling	19.83 (0.64)	0.92 (0.00)
WMT-16	1	Poisson sampling	2.35 (0.07)	0.84 (0.00)
BSD	∞	Random shuffling	10.09 (2.75)	0.90 (0.01)
BSD	1000	Random shuffling	1.36 (0.67)	0.87 (0.01)
BSD	1000	Poisson sampling	1.01 (0.07)	0.87 (0.00)
BSD	5	Random shuffling	0.06 (0.05)	0.85 (0.01)
BSD	5	Poisson sampling	0.06 (0.06)	0.84 (0.02)
BSD	1	Random shuffling	0.00 (0.01)	0.45 (0.22)
BSD	1	Poisson sampling	0.00 (0.00)	0.65 (0.15)
ClinSPEn-CC	∞	Random shuffling	5.42 (2.41)	0.86 (0.02)
ClinSPEn-CC	1000	Random shuffling	0.03 (0.02)	0.75 (0.01)
ClinSPEn-CC	1000	Poisson sampling	0.70 (0.19)	0.78 (0.00)
ClinSPEn-CC	5	Random shuffling	0.80 (0.56)	0.79 (0.00)
ClinSPEn-CC	5	Poisson sampling	0.83 (0.27)	0.79 (0.00)
ClinSPEn-CC	1	Random shuffling	0.50 (0.20)	0.78 (0.00)
ClinSPEn-CC	1	Poisson sampling	0.54 (0.22)	0.78 (0.00)

Tabella di tipo 1

Model	L=0.5	L=0.2	L=0.1	L=0.05	L=0.02
GPT2-small	0.131	0.135	0.131	0.135	0.132
GPT2-small-share-encoder	**0.248**	**0.265**	**0.264**	**0.255**	**0.251**

Tabella di tipo 2

	MOS	
TTS Model	**Without Stress**	**With Stress**
Pitch	**4.25**	3.95
Pitch and Energy	**4.36**	4.21

Tabella di tipo 3

CLAIMS EXTRACTION - TIPO 1

Lo script che abbiamo utilizzato per l'estrazione svolge i seguenti passaggi:

- Lettura della tabella HTML.
- Identificazione delle metriche o delle specifiche con API Gemini.
- Creazione di claims nel formato |{specifications}, {measure}, {outcome}|.
- Salvataggio in file JSON.

Dataset	Lang. Pair	# Trn.+Vld.	# Test
WMT-16	DE-EN	4,551,054	2,999
BSD	JA-EN	22,051	2,120
ClinSPEn-CC	ES-EN	1,065	2,870


```
[
  {
    "Claim 0": "|{|Dataset, WMT-16|,|Lang. Pair, DE-EN|}, # Trn.+Vld., 4,551,054|"
  },
  {
    "Claim 1": "|{|Dataset, WMT-16|,|Lang. Pair, DE-EN|}, # Test, 2,999|"
  },
  {
    "Claim 2": "|{|Dataset, BSD|,|Lang. Pair, JA-EN|}, # Trn.+Vld., 22,051|"
  },
  {
    "Claim 3": "|{|Dataset, BSD|,|Lang. Pair, JA-EN|}, # Test, 2,120|"
  },
  {
    "Claim 4": "|{|Dataset, ClinSPEn-CC|,|Lang. Pair, ES-EN|}, # Trn.+Vld., 1,065|"
  },
  {
    "Claim 5": "|{|Dataset, ClinSPEn-CC|,|Lang. Pair, ES-EN|}, # Test, 2,870|"
  }
]
```

CLAIMS EXTRACTION - TIPO 2

Lo script che abbiamo utilizzato per l'estrazione svolge i seguenti passaggi:

- Lettura della tabella HTML.
- Le intestazioni sono prese dalla prima riga della tabella. Le righe successive contengono specifiche e valori.
- Utilizzo di Gemini per identificare la metrica ed estrae le specifiche basandosi su caption e paragrafi.
- Creazione di claims nel formato |{specifications}, {measure}, {outcome}|.
- Salvataggio in file JSON.

Table 1: The delay improvement performance of using the shared encoder as the branch predictor on the en→vi direction wait-k [3] method.

Model	K=1	K=3	K=5	K=7	K=9
GPT2-small	0.157	0.179	0.131	0.176	0.166
GPT2-small-share-encoder	**0.210**	**0.225**	**0.248**	**0.211**	**0.198**


```
[
  {
    "Claim 0": "|{|Model, GPT2-small|,|wait-k, K=1|}, delay improvement performance, 0.157|"
  },
  {
    "Claim 1": "|{|Model, GPT2-small|,|wait-k, K=3|}, delay improvement performance, 0.179|"
  },
  {
    "Claim 2": "|{|Model, GPT2-small|,|wait-k, K=5|}, delay improvement performance, 0.131|"
  },
  {
    "Claim 3": "|{|Model, GPT2-small|,|wait-k, K=7|}, delay improvement performance, 0.176|"
  },
 {…}
]
```

CLAIMS EXTRACTION - TIPO 3

Lo script che abbiamo utilizzato per l'estrazione svolge i seguenti passaggi:

- Lettura della tabella HTML.
- Estrazione delle specifiche con attributo '*Colspan*'.
- Header estratto identificando il campo '*Colspan*' e valori processati dalla terza riga.
- Utilizzo di Gemini per identificare la metrica ed estrae le specifiche basandosi su caption e paragrafi.
- Salvataggio in file JSON.

		Translation Quality		User Effort		
Model	**Language Pair**	**TER [↓]**	**BLEU [↑]**	**WSR [↓]**	**KSR [↓]**	**MAR [↓]**
mT5	De–En	69.3	15.1	52.17	64.84	19.78
	En–De	74.2	13.2	63.64	66.00	17.00
	Es–En	65.3	18.1	44.06	51.49	14.40
	En–Es	64.3	18.4	46.45	55.57	13.92
	Fr–En	66.3	18.6	44.74	52.62	15.02
	En–Fr	81.8	17.8	48.34	55.73	15.16
mBART	De–En	52.4	29.7	52.17	68.13	19.78
	En–De	57.0	27.1	50.00	56.00	14.00
	Es–En	52.1	30.5	33.08	38.95	12.19
	En–Es	48.2	33.3	34.41	41.09	11.68
	Fr–En	48.4	33.6	32.35	37.90	12.35
	En–Fr	56.0	39.1	29.92	34.38	11.07


```
[
  {
    "Claim 0": "|{|Model, mT5|,|Language Pair, De-En|,|Translation Quality, TER↓}, TER↓, 69.3|"
  },
  {
    "Claim 1": "|{|Model, mT5|,|Language Pair, De-En|,|Translation Quality, BLEU↓}, TER↓, 15.1|"
  },
  {
    "Claim 2": "|{|Model, mT5|,|Language Pair, De-En|,|User Effort, WSR↓}, WSR↓, 52.17|"
  },
  {
    "Claim 3": "|{|Model, mT5|,|Language Pair, De-En|,|User Effort, KSR↓}, KSR↓, 64.84|"
  },
  {…},
]
```

GROUND-TRUTH

L’obiettivo principale è stato quello di verificare se l’estrazione svolta con il sistema andasse ad individuare la maggior parte se non tutti i claim contenuti nelle tabelle.
E’ stata svolta un’analisi a mano, considerando tutte le tabelle e definendo delle metriche per valutare la similarità.
La similarità tra due claim è stabilita in funzione di una soglia scelta del 90%.
Si può notare che i valori sono gli stessi, perché c’è una corrispondenza completa tra il set degli elementi estratti e il set della groundtruth.

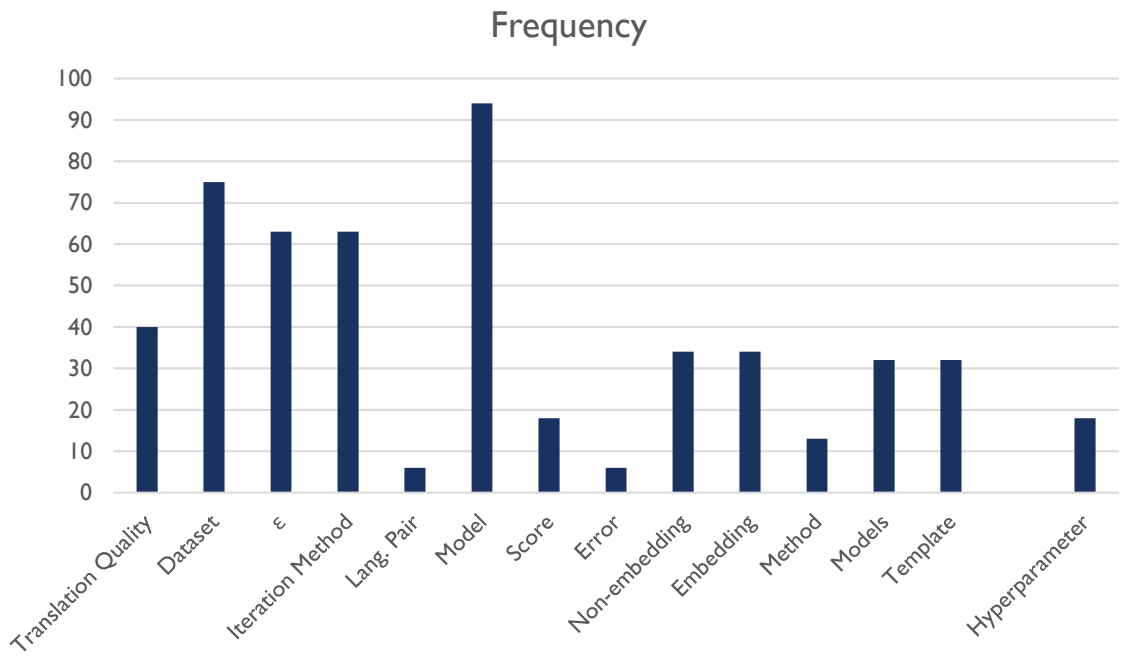
Precision	Recall	F1-Score
0.72	0.72	0.72

PROFILING

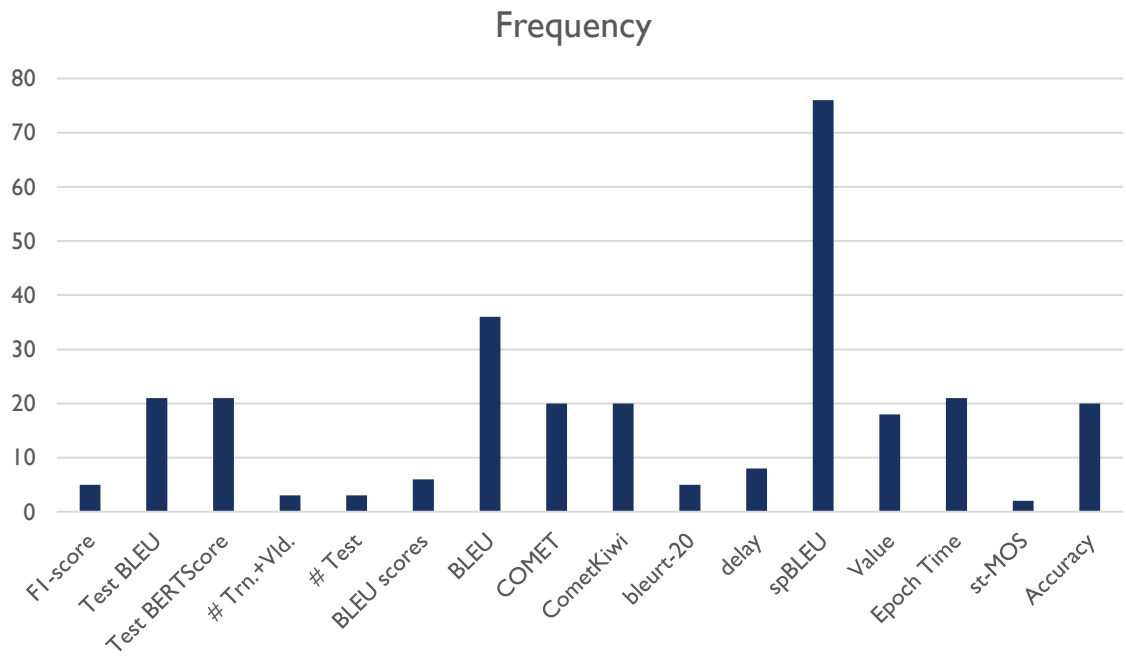
Questo task analizza le affermazioni estratte per generare dati di profilazione completi. Fornisce approfondimenti statistici calcolando distribuzioni e medie. Le distribuzioni estratte sono le seguenti:

- Distribuzione dei nomi delle specifiche.
- Distribuzione delle metriche.
- Distribuzione dei valori per ogni specifica.
- Valori medi associati ad ogni metrica.

PROFILING

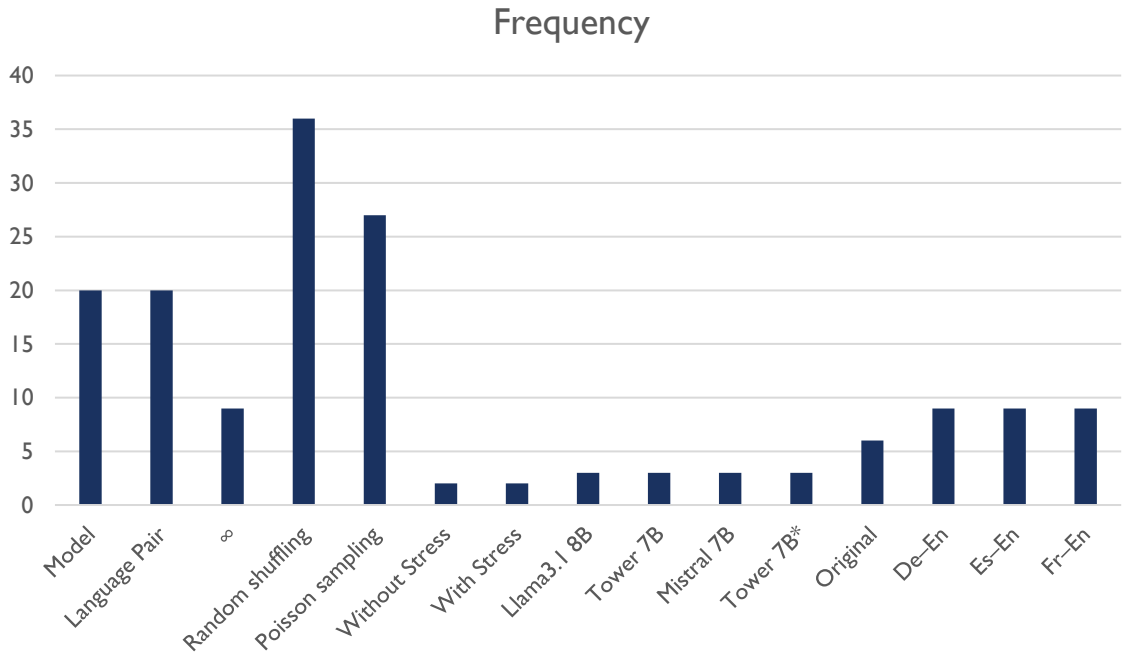


Specifiche

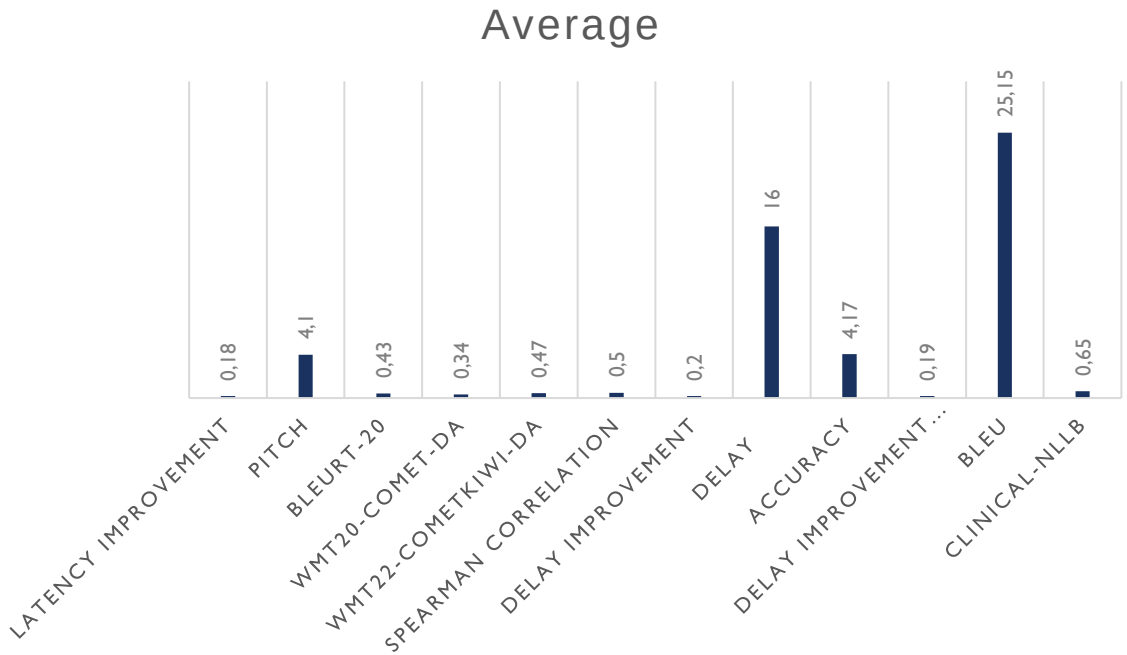


Metriche

PROFILING



Valori specifiche



Valori medi delle metriche

ALIGNMENT

Questo task standardizza e unifica la terminologia nelle affermazioni estratte, garantendo coerenza tra metriche e specifiche.

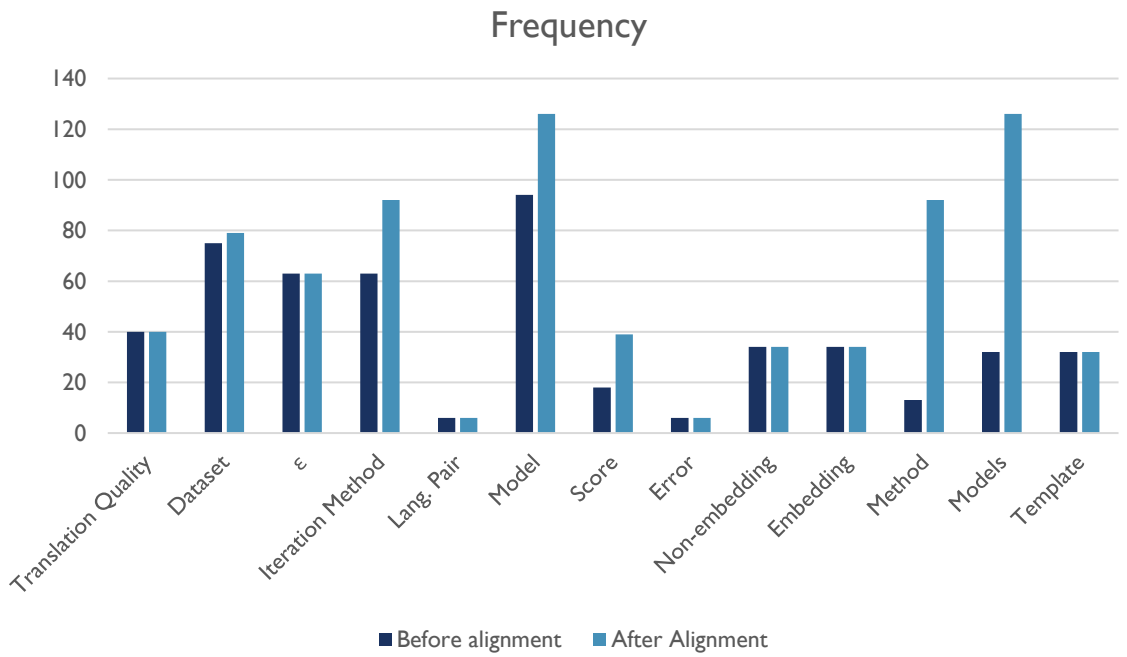
E’ stato inizialmente creato un dizionario dei sinonimi, utilizzando un modello pre-addestrato di tipo SentenceTransformer, nello specifico **all-MiniLM-L6-v2.**

Il processo di costruzione del dizionario è composto dalle seguenti fasi:

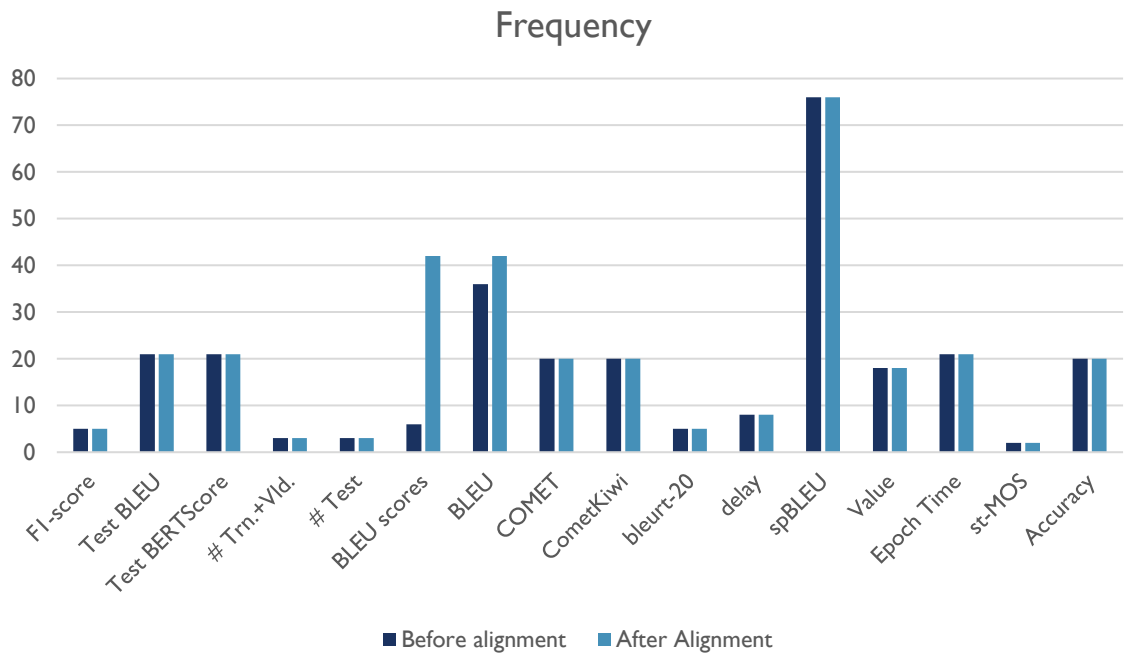
- **Normalizzazione dei termini.**
- **Calcolo degli embeddings.**
- **Calcolo matrice similarità.**
- **Clusterizzazione dei sinonimi.**
- **Gestione dei duplicati con risultato finale.**

```
"lang. pair": ["lang. pair", "language pair"],
"score": ["score", "diff. in scores", "value"],
"dataset": ["dataset", "data set"],
"model": ["model", "models", "models, gpt-
    4o","models, gpt-4"],
"method": ["iteration method", "method", "methods"]
```

PROFILING – WITH ALIGNMENT

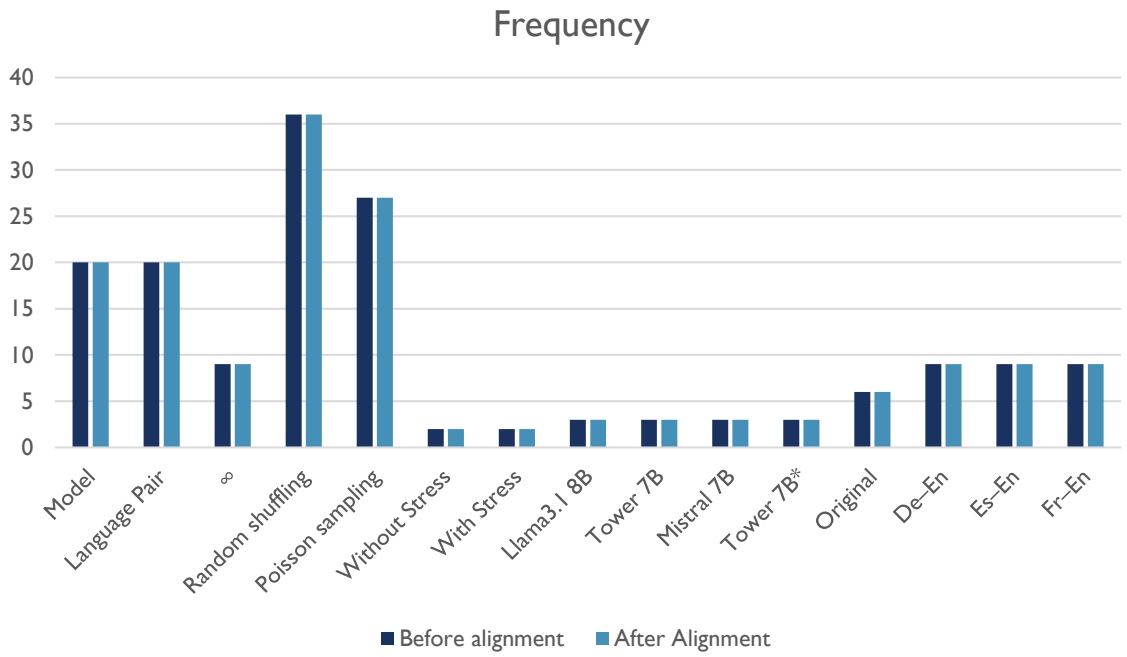


Specifiche

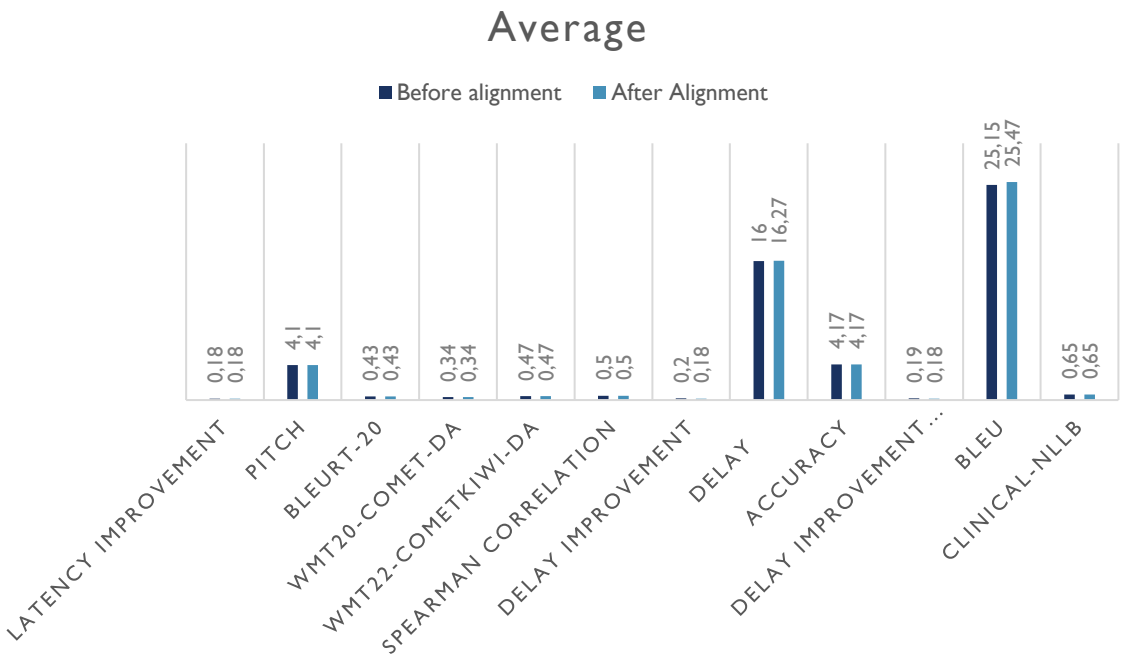


Metriche

PROFILING – WITH ALIGNMENT



Valori specifiche



Valori medi delle metriche

ALIGNMENT

Il processo vero e proprio si occupa di caricare un file JSON contenente i campi e i valori allineati (a partire dai claims estratti) e di unificarli in un'unica rappresentazione, prendendo in considerazione anche il dizionario dei sinonimi.

Si compone delle seguenti fasi:

- **Normalizzazione dei termini.**
- **Caricamento del dizionario dei sinonimi.**
- **Identificazione dei sinonimi per ogni campo.**
- **Unione dei campi simili, usando il dizionario dei sinonimi.**
- **Creazione di un nuovo dataset unificato.**
- **Salvataggio del file risultante.**


```
"aligned_names": {
"dataset": [
        "2311.14465_3_claims_38_1",
        "2311.14465_3_claims_29_1",
        "2410.07830_1_claims_8_1",
        "2311.14465_3_claims_32_1",
        "2409.17939_2_claims_1_1"  ],
"mos": [
        "2403.04178_1_claims_0_1",
        "2403.04178_1_claims_1_1",
        "2403.04178_1_claims_2_1",
        "2403.04178_1_claims_3_1"   ],
"method": [
        "2409.17939_1_claims_12_1",
        "2410.06338_3_claims_6_1",
        "2409.17939_1_claims_9_1",
        "2410.06338_4_claims_1_1",
        "2409.17939_1_claims_4_1",
        "2311.14465_3_claims_31_3"  ],
```

CONCLUSIONI E SFIDE INCONTRATE

Dati rumorosi

Complessità semantica

Limitazioni computazionali

SVILUPPI FUTURI

Estensione a nuovi tipi di dati

Miglioramento dell’accuratezza dell’estrazione

Automatizzazione completa

GRAZIE PER L'ATTENZIONE!