

Identificazione e disambiguazione della costruzione NPN con BERT

Caso studio di Scivetti e Schneider sulla costruzione inglese

Corso di Semantica a.a. 2025/2026

18 Novembre 2025

Università di Bologna

La costruzione NPN

Il Dataset

Training e Test set

BERT, modello di encoder

Task 1: Identificazione

Task 2: Identificazione (perturbando l'ordine delle parole)

Task 3: Disambiguazione semantica

La costruzione NPN

Cos'è la costruzione NPN

La costruzione NPN è stata ampiamente studiata in inglese da una prospettiva costruzionista, in particolare da Jackendoff (2008) e Sommerer & Baumann (2021).

Construction Schema

*Noun*₁ Preposition *Noun*₂

I due nomi nella costruzione devono avere forma identica, anche per la declinazione del numero. La costruzione non ammette nomi accompagnati da determinati.

Cos'è la costruzione NPN

La costruzione può comparire in diverse posizioni sintattiche, ad esempio come modificatore avverbiale o come modificatore nominale.

Esempi

- *I need you to get this **word for word**.*
(modificatore avverbiale)
- *There is a rebellious quality to your **day to day** responses which have not gone unnoticed.* (modificatore nominale)

Significati delle costruzioni NPN istanziate da *to*

SUCCESSION

Testo del blocco esplicativo.

JUXTAPOSITION

Testo del blocco esplicativo.

II Dataset

- estrazione da COCA
- Eliminazione dei casi PNPN
- Identificazione dei distrattori
- Annotazione per tutte le istanze delle etichette semantiche

Doppia annotazione

- Annotato **25%** del dataset
- Accordo grezzo: **84%**
- Cohen's kappa: **0.754** (accordo forte)

Dimensioni del dataset

- **6599** istanze totali (N-to-N)
- **1885** istanze con doppia annotazione

Training e Test set

Near Minimal Pairs e Distrattori

NtoN distractors

Oltre alle reali istanze della costruzione NtoN, il corpus contiene anche pattern superficiali **Noun + to + Noun** che non sono costruzioni NtoN. Derivano da contesti sintattici diversi (es. verbi che reggono un oggetto e una PP con *to*): *stick plastic to plastic*, *time to time travel*, ecc.

Questi casi non esprimono il significato della costruzione ma forniscono utili **esempi negativi** per testare se il modello.

Near minimal pairs

Poiché condividono la stessa forma superficiale dei veri NtoN, questi distrattori costituiscono **near minimal pairs**: frasi grammaticali, naturali, quasi identiche in superficie, ma con **struttura e significato diversi**.

Dataset

Nel case study **456** near minimal pairs come distrattori dal corpus.

Split training test set

Evitare overfitting

Max 20 occorrenze per lemma per evitare overfitting e ridurre la sproporzione tra lemmi altamente frequenti e lemmi rari.

Controllo della generalizzazione

Generazione split casuali di train/test basati sul lemma del nome presente nella costruzione NtoN, in modo che nessun lemma compaia sia nel training set sia nel test set.

Bilanciamento training

Poiché il numero di distrattori è significativamente inferiore rispetto alle istanze della costruzione, per bilanciare le categorie nel training set è stato utilizzato l'80% dei distrattori, abbinato allo stesso numero di costruzioni. Il test set è quindi composto dal restante 20% dei distrattori, insieme a tutte le costruzioni eccedenti quelle usate per il training.

BERT, modello di encoder

Embedding

Le parole e le frasi vengono trasformate in **vettori numerici** chiamati embedding. Questi vettori catturano somiglianze semantiche e relazioni tra parole, permettendo al modello di “comprendere” il testo.

Perché trasformare il testo

Il testo deve diventare numerico per essere elaborato dai modelli. Trasformare significa codificare ogni parola in uno spazio continuo dove vicinanza = somiglianza.

Transformer

Modello basato su **self-attention**: ogni parola osserva tutte le altre per capire il contesto. Non usa ricorrenza e permette di gestire sequenze lunghe in parallelo.

Encoder

L'encoder produce rappresentazioni contestuali di ogni parola, che riflettono sia il significato intrinseco sia le relazioni con le altre parole nel testo.

BERT: bidirezionale e contestuale

Cos'è BERT

BERT è un Transformer **solo encoder**, che legge il contesto a sinistra e a destra di ogni parola. Produce embedding contestuali che catturano significato, struttura sintattica e relazioni tra parole.

Addestramento

- **Masked Language Modeling**: predire token mascherati.
- **Next Sentence Prediction**: capire se due frasi sono in sequenza.

Perché usarlo

Le rappresentazioni di BERT possono essere adattate a molti task NLP: classificazione, NER, question answering. Per questo caso studio, mostrano come la semantica delle costruzioni è rappresentata.

Task 1: Identificazione

Task 1: Identificazione

Definizione del task

Distinguere le istanze autentiche della costruzione NtoN dagli esempi autentici del corrispondente pattern distrattore.

Control classifier

Le etichette vengono randomizzate e assegnate in modo deterministico al word type.

Le performance dovrebbero attestarsi near chance.

Non-contextual baseline (GloVe)

Valuta la performance basata solo su informazioni lessicali, senza contesto.

Dovrebbero attndersi, in virtù dei campi semnatici ricorrenti nella costruzione NtoN come espressioni temporali e parti del corpo, performance non trascurabili.

A cosa serve la baseline?

Control classifier

Informa sulla bontà del classificatore

Le performance dovrebbero attestarsi near chance.

Non-contextual baseline (GloVe)

Tutto ciò che supera questa performance potrebbe essere attribuito alle informazioni aggiuntive catturate da BERT attraverso il significato contestuale.

Task 2: Identificazione (perturbando l'ordine delle parole)

Esempio di blocco

Testo del blocco esplicativo.

Task 3: Disambiguazione semantica

Esempio di blocco

Testo del blocco esplicativo.