La piattaforma T2K: dal testo alla conoscenza

Felice Dell'Orletta

ItaliaNLP Lab – www.italianlp.it

Istituto di Linguistica Computazionale «A. Zampolli»

16 ottobre 2014

Italian Natural Language Processing Laboratory @ILC

Persone

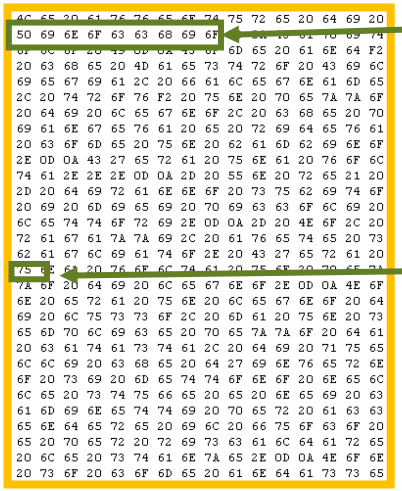
Simonetta Montemagni
Felice Dell'Orletta
Giulia Venturi
Andrea Cimino
Dominique Brunato

www.italianlp.it

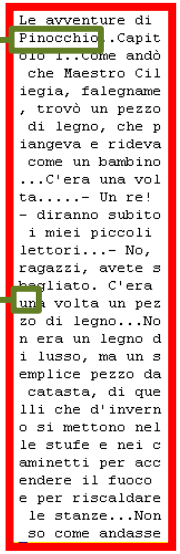
ItaliaNLP Lab

- **Obiettivi**
  - sviluppo di tecnologie linguistiche allo scopo di estrarre ed organizzare il contenuto (sia linguistico che di conoscenza) nascosto nei testi
- **Principali linee di ricerca**
  - **analisi linguistica automatica dei testi:**
    - sviluppo di strumenti multi-lingua per l'analisi linguistica multi-livello del testo
    - costruzione di corpora per l'addestramento e la valutazione di algoritmi basati su metodi di apprendimento automatico
    - sviluppo di metodi per adattare strumenti di NLP a domini specifici e varietà di lingue non canoniche
  - **estrazione di conoscenza:**
    - estrazione ed organizzazione di terminologia di dominio
    - annotazione semantica di entità nominate ed entità rilevanti per uno specifico dominio
    - estrazione di relazioni tra le entità estratte
    - studio dei modelli di variazione linguistica: ricostruzione del profilo linguistico dei testi rispetto al dominio, il genere testuale ed il registro; studio delle variazioni dialettali e sociolinguistiche
  - **prototipi software**
    - ***LinguA: linguistic annotation pipeline*.** Catena di analisi linguistica in linea con lo stato dell'arte che combina sia sistemi a regole che algoritmi basati sull'apprendimento automatico
    - ***Text-to-Knowledge (T2K).*** Piattaforma software per l'estrazione e organizzazione della conoscenza linguistica e di dominio dai testi
    - ***READ-IT: Assessing Readability of italian Text.*** READ-IT è il primo sistema avanzato per l'analisi della leggibilità dei testi scritti in lingua italiana

Le tecnologie del linguaggio: perché?



Non tutti guardano le cose allo stesso modo

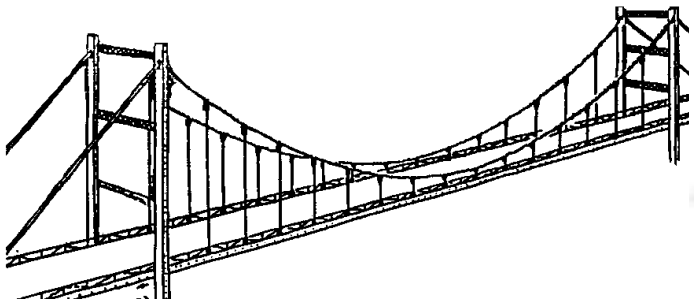


Le tecnologie del linguaggio: cosa sono?

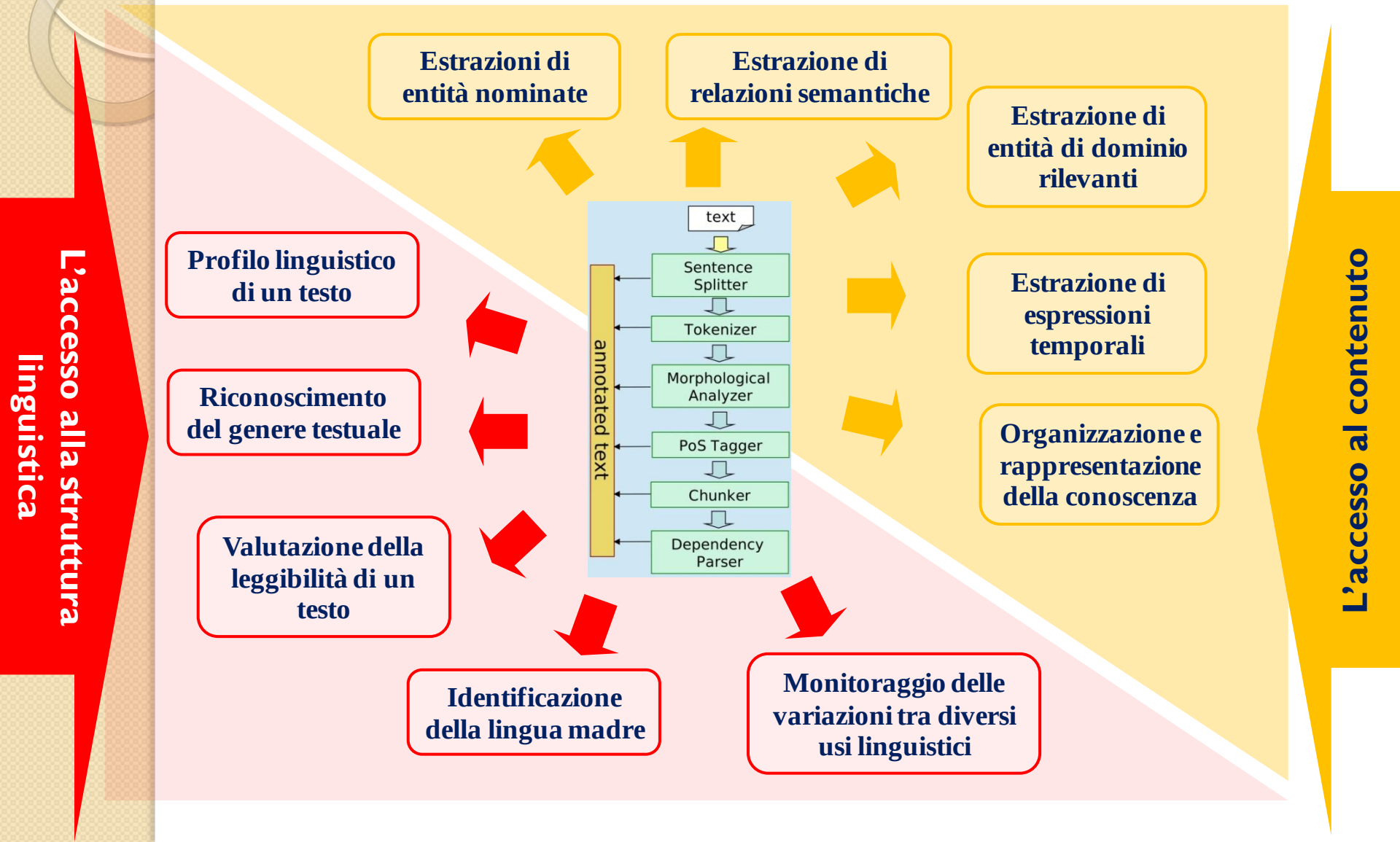
- Sistemi in grado di accedere al contenuto informativo dei testi attraverso l'elaborazione automatica del linguaggio (*Natural Language Processing*)

Un 'ponte' tra il testo e il contenuto

- Conoscenza linguistica
  - morfo-sintattico, sintattico, semantico-lessicale
- Conoscenza di dominio



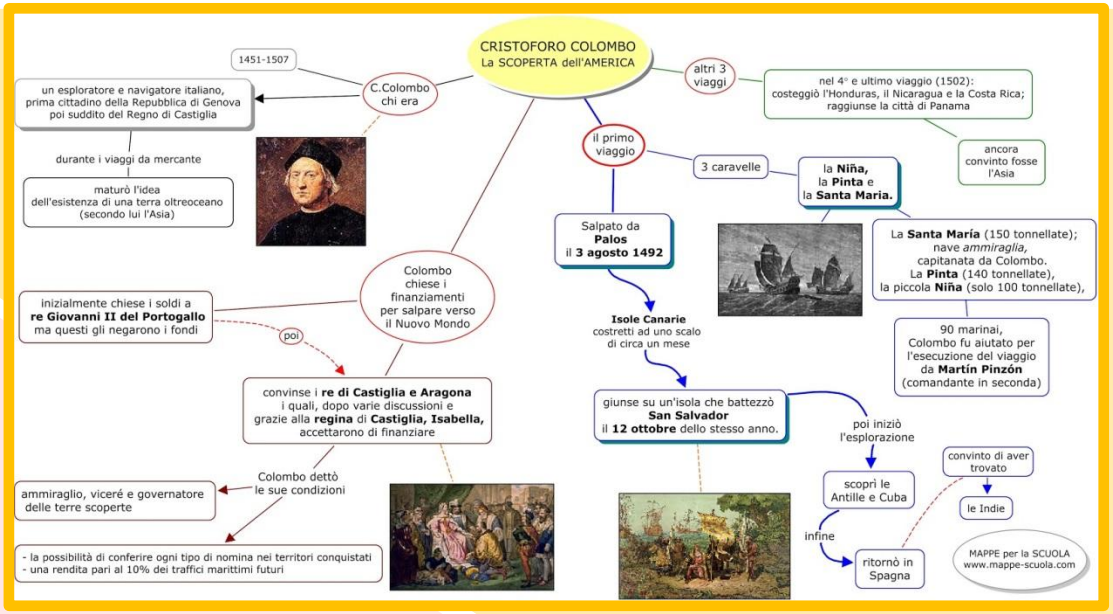
ItaliaNLP Lab: tecnologie del linguaggio



Le tecnologie del linguaggio per …

L'accesso alla struttura linguistica

Costruzione di mappe concettuali dai testi

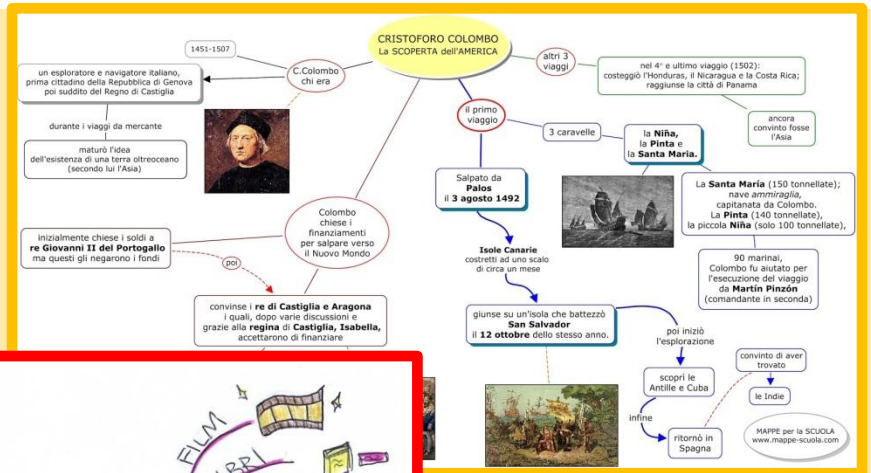


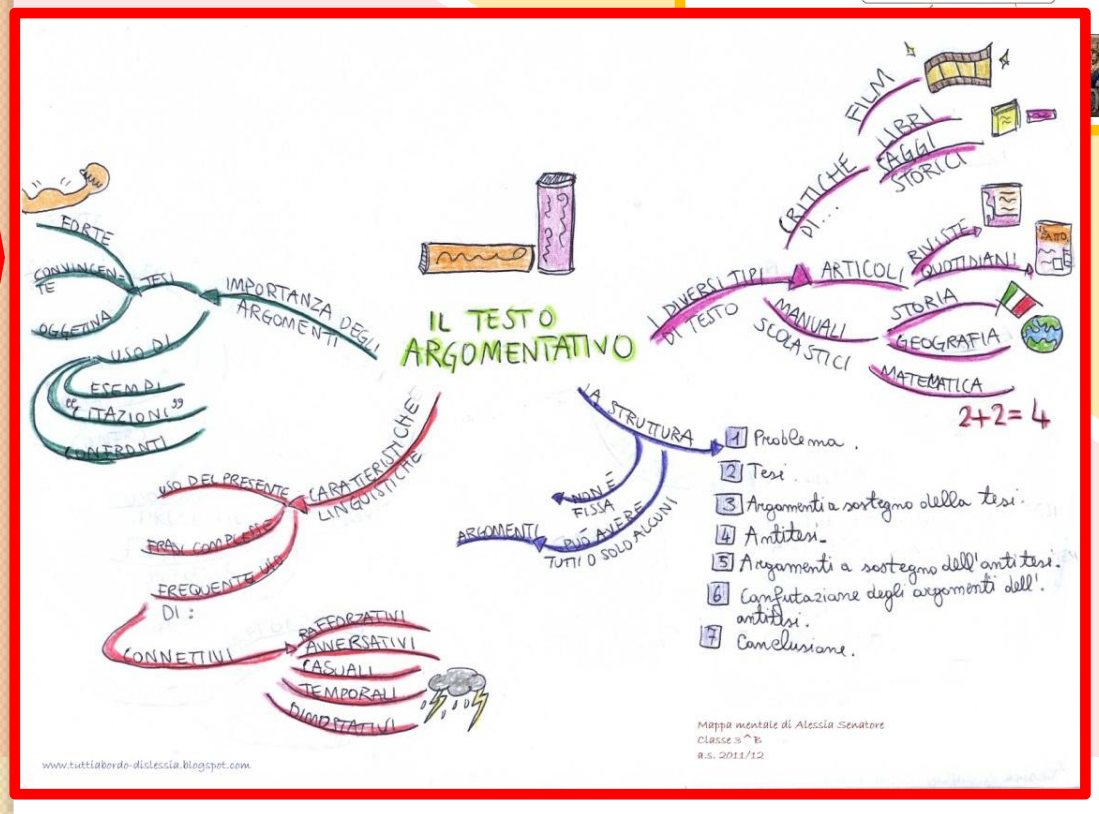
L'accesso al contenuto

Le tecnologie del linguaggio per …

L'accesso alla struttura linguistica

Costruzione di mappe concettuali dai testi

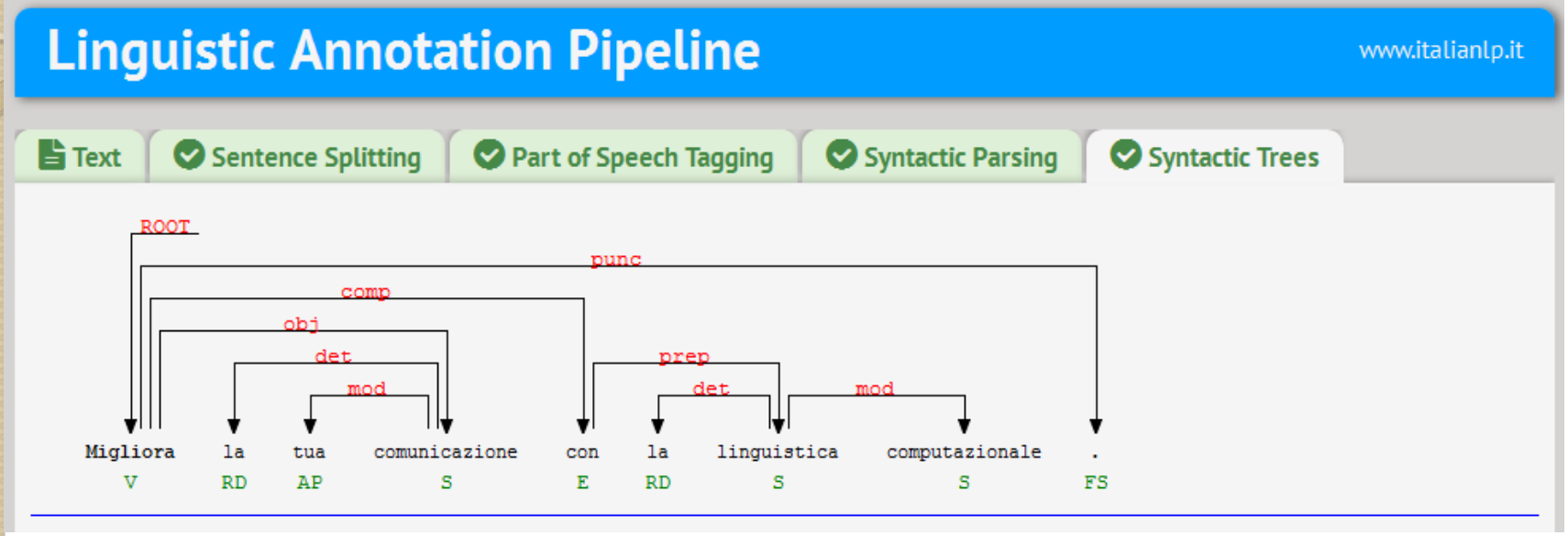




Analisi e verifica delle caratteristiche linguistiche dei testi

L'accesso al contenuto

LinguA: Linguistic Annotation pipeline



- Catena di analisi linguistica
  - Sviluppata da ILC e Università di Pisa
  - http://www.italianlp.it/demo/linguistic-annotation-tool/

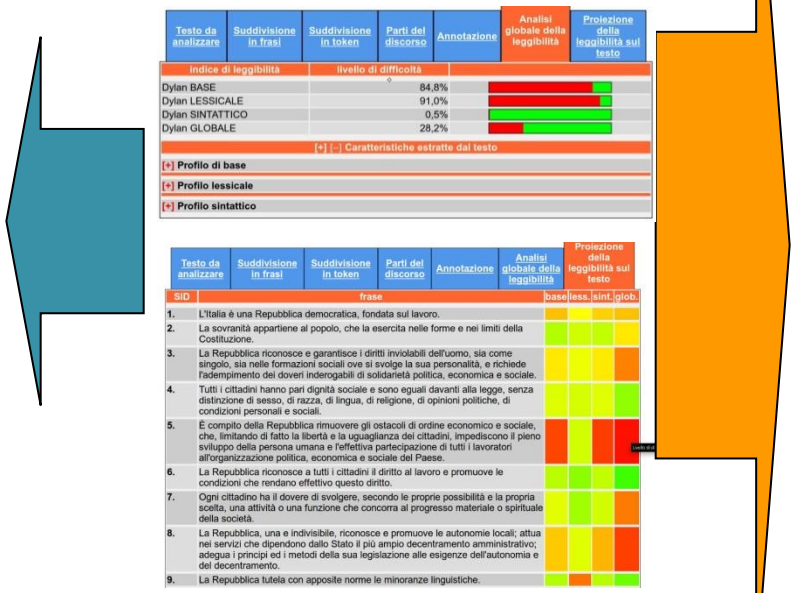
READ-IT: Assessing Readability of Italian Texts

READ–IT combina caratteristiche tradizionali estratte dal testo con informazioni morfo-sintattiche e sintattiche. READ–IT valuta la leggibilità sia rispetto all'intero documento che alle singole frasi, supportando la semplificazione del testo rispetto allo specifico audience obiettivo. *www.italianlp.it/demo/*

Monitoraggio delle caratteristiche linguistiche di collezioni di testi

Studio dei fattori che rendono un testo complesso

Modelli della comprensione linguistica



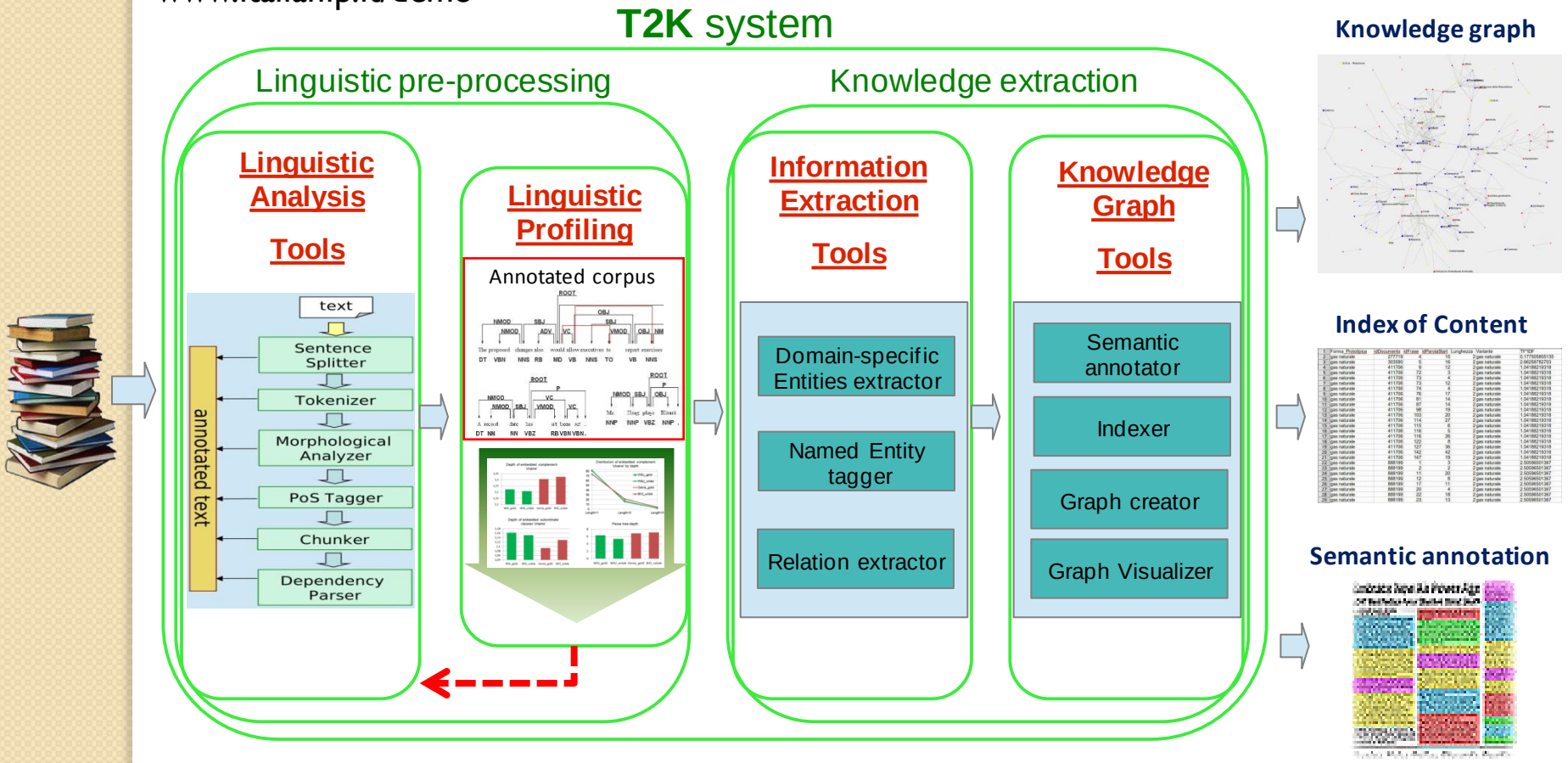
Valutazione dell'**efficacia comunicativa** di testi nella comunicazione

- **Insegnante-studente** (Progetto CNR "Migrazioni")
- **Amministratore-Cittadino** (Osservatorio per la redazione di atti amministrativi – Crusca – ITTIG-CNR)
- **Operatore di Call Center-utente finale** (collaborazione con Vodafone)
- **Medico-Paziente** (progetto SUIT-HEART Progetto Italiano "Istituto Toscano Tumori")
- **Autore editoria scolastica-studenti** (progetto Regione Toscana iSLe, in corso)

READ-IT: assessing readability of Italian texts with a view to text simplification. Dell'Orletta F., Montemagni S., Venturi G. In Proceedings of SLPAT 2011 , PA, USA.

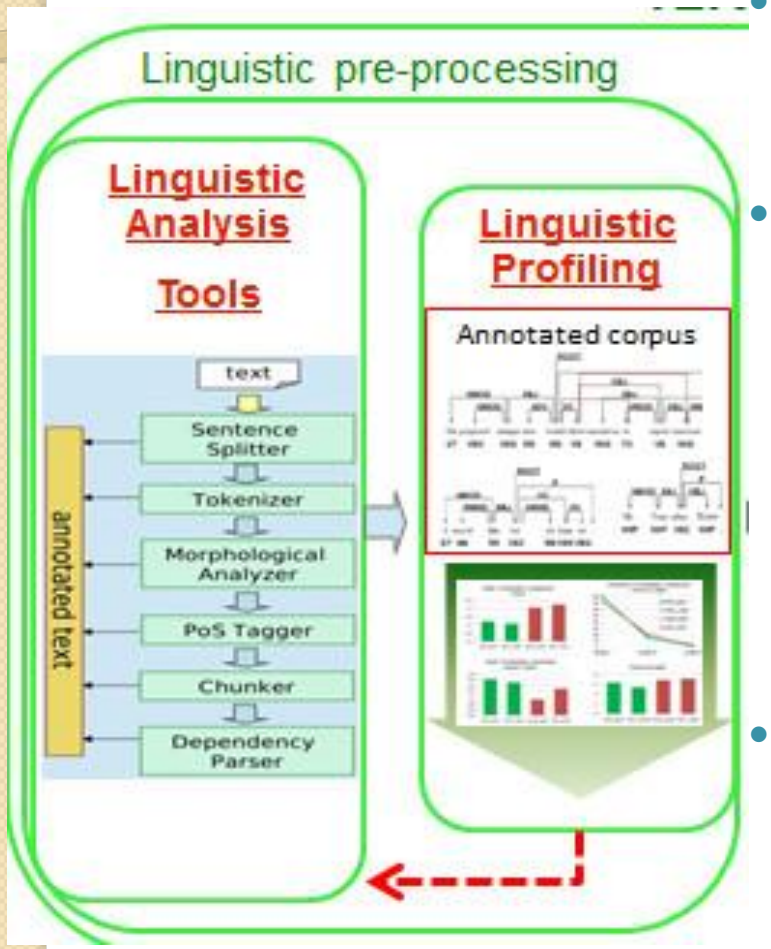
From Text to Knowledge:T2K

T2K combina una batteria di tool NLP, tecniche di analisi statistica e di machine learning, allo scopo di produrre una rappresentazione strutturata dell'informazione estratta a partire da documenti non strutturati. www.italianlp.it/demo



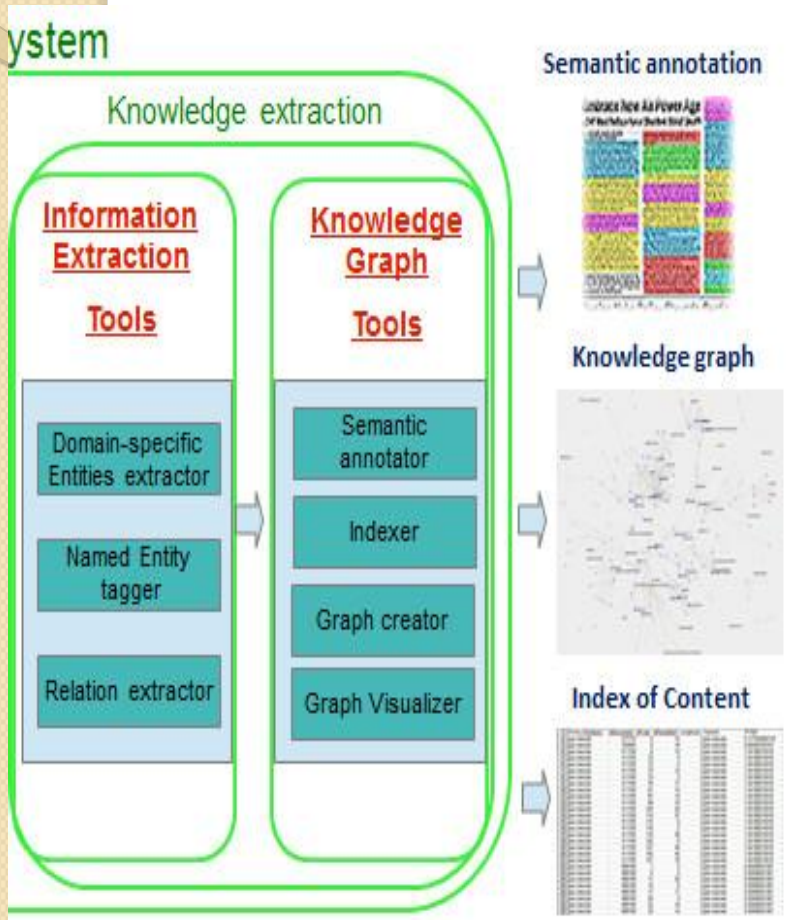
T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts. Dell'Orletta F., Venturi G., Cimino A., Montemagni S. In Proceedings of LREC 2014, 26-31 May, Reykjavik, Iceland.

Estrazione di conoscenza linguistica



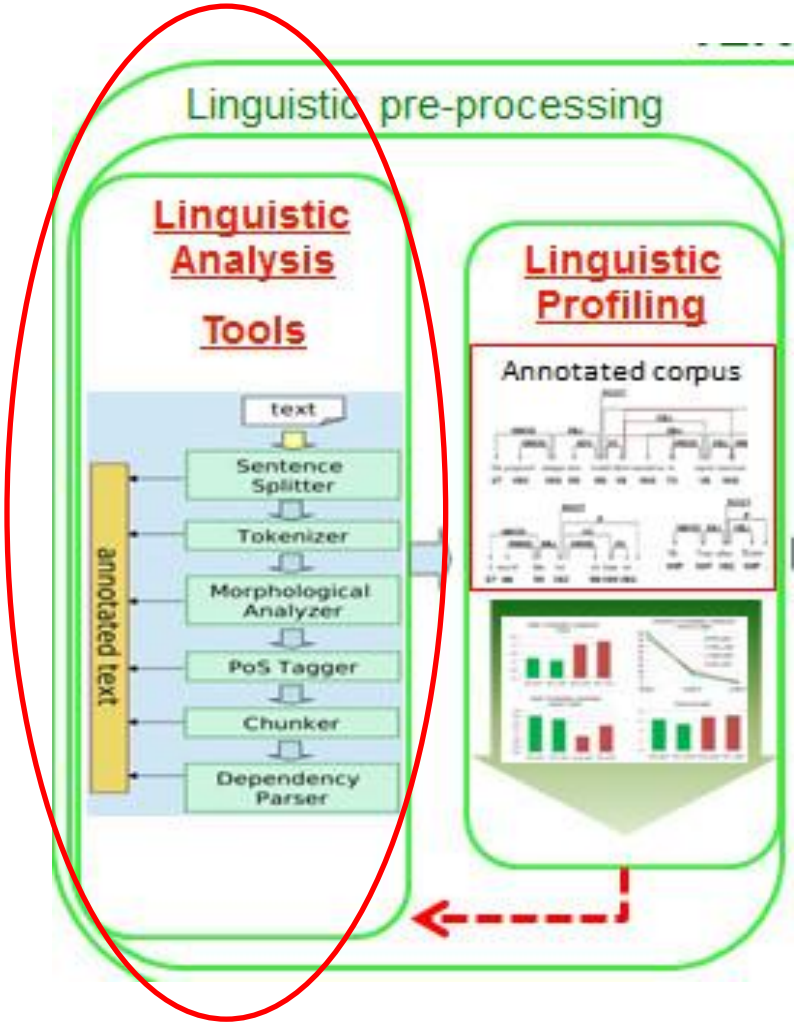
- The linguistically analyzed corpus is used by the **linguistic profiling module** to investigate the *form* of a text rather the *content*
- The distribution of a wide range of linguistic features (lexical, morpho-syntactic and syntactic) is aimed at
  - assessing the readability level (Dell'Orletta et al., 2011)
  - native language identification (Cimino et al., 2012)
  - determining the text genre (Dell'Orletta et al., 2013)
- Moreover, they can be used to refine the construction of the corpus
  - In terms of homogeneity and representativeness of a given domain

Estrazione di informazione di dominio



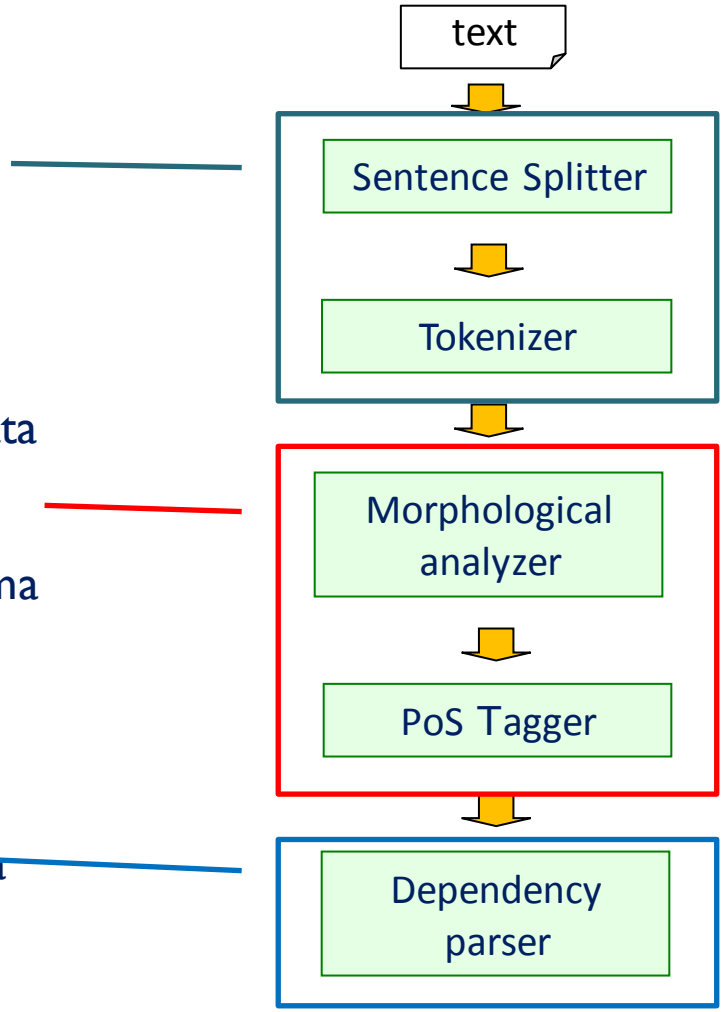
- The IE tools allow extracting
  - domain-specific entities (Bonin et al. 2010)
    - e.g. nominal terminology, verbs (both single- and multi-word expressions)
  - Named entities
    - i.e. Person, Location, Organization and Geopolitical
  - relations between the extracted entities
    - taxonomical
      - e.g. health research, international research, cancer research or research projects, research infrastructure
    - co-occurence within the same context and similarity on the basis of shared contexts
- They result in
  - multi-dimensional knowledge representation graph
  - document collection index and semantic annotation

Catena di analisi linguistica



Catena di analisi linguistica

- **Segmentazione in frasi e tokenizzazione** (ovvero segmentazione del testo in parole ortografiche o tokens)
- **Annotazione morfo-sintattica**
  - a ogni token del testo viene associata informazione relativa alla categoria grammaticale che la parola ha nel contesto specifico e il relativo lemma
- **Annotazione sintattica a dipendenze**
  - analisi della struttura sintattica della frase in termini di relazioni di dipendenza (es. soggetto, oggetto, etc.)



Annotazione linguistica automatica: requisiti di base

- **robustezza** nel trattare input mal formato o non conforme alle regole generali della lingua italiana
- **accuratezza** dei risultati prodotti
- **efficienza** nella capacità di gestire ingenti quantità di dati
- **adattabilità** a diversi domini, generi testuali, registri linguistici così come a nuove lingue

Le "insidie" del linguaggio: alcuni esempi

Nome o verbo?

La vecchia **porta** la sbarra

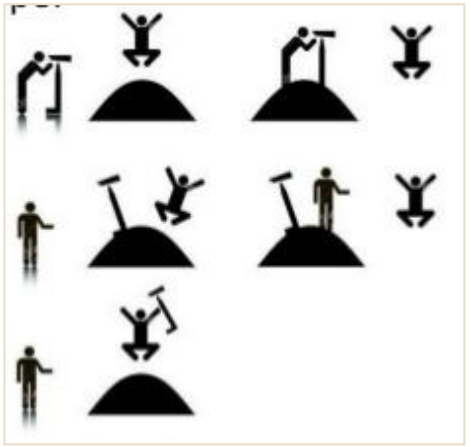
Quale senso di *interesse*?

Il tasso di **interesse** è variabile anche in funzione della moneta di riferimento

Ha mostrato molto **interesse** per la Linguistica Computazionale

Ho visto l'uomo **sulla collina *con il telescopio***

Chi è sulla collina?
Chi ha il telescopio?



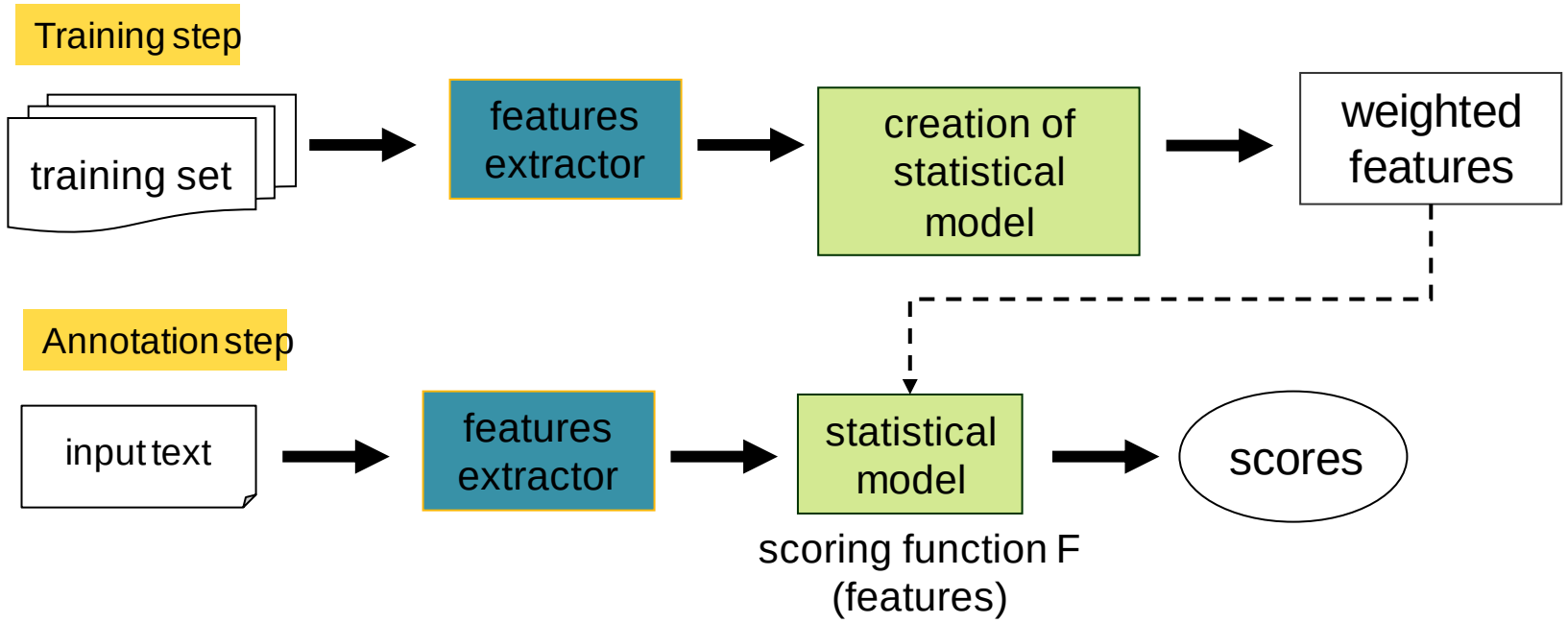
Annotazione linguistica stocastica

- Lo stato dell'arte dell'annotazione linguistica è rappresentato da sistemi basati su algoritmi di appredimento automatico
  - molto efficienti
  - estremamente accurati nella risoluzione di problemi di classificazione
- Annotazione linguistica come classificazione statistica
  - non esiste una metodologia standard per eseguire tale trasformazione, dipende dal tipo di compito che dobbiamo affrontare
  - questa trasformazione è più semplice per compiti che coinvolgono un unico token per volta (ad esempio l'analisi morfo-sintattica) mentre è più complessa in compiti nei quali devono essere identificate delle relazioni tra più elementi della frase come ad esempio l'analisi sintattica.

Annotazione linguistica stocastica

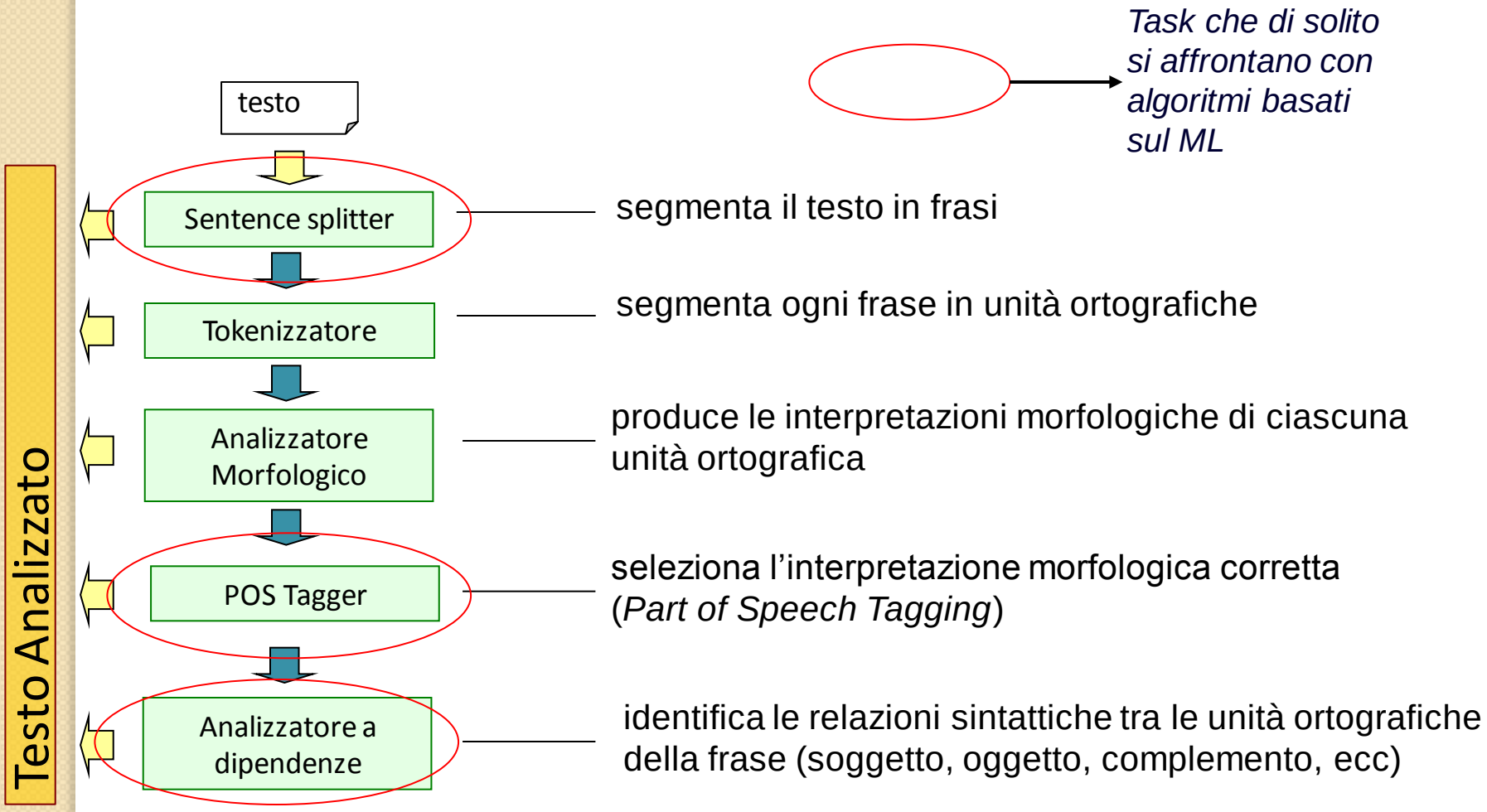
- Indipendentemente dall'algoritmo di apprendimento utilizzato sono richiesti tre ingredienti fondamentali per la creazione del modello statistico:
  - l'**insieme delle categorie linguistiche** da assegnare
  - il **corpus di addestramento** (ovvero un insieme di esempi pre-annotati classificati a mano)
  - un **insieme di tratti descrittivi**, accuratamente selezionati sulla base del compito di classificazione da svolgere
- A partire da un corpus di addestramento viene costruito un modello statistico per l'annotazione linguistica del testo.
- Il modello statistico viene utilizzato in fase di analisi di nuovi testi.

Annotazione linguistica stocastica



- 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 viene poi applicato a esempi sconosciuti per assegnare loro la classe più probabile, dato il modello e l'insieme di tratti pertinenti.
- **scoring function**: usa sia le "weighted features" sia le "extracted features" per identificare la classificazione più probabile

ML nei compiti di annotazione linguistica



Sentence Splitting

- Il primo passo dell'analisi linguistica di base è il "*sentence splitting*": l'identificazione delle frasi all'interno del testo.
- Il modulo utilizza algoritmi basati sul ML per la classificazione dei punti in 2 classi: "fine frase" e "abbreviazione"

Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato.

- Il danno non poteva essere sottovalutato.
- Il sig. Rossi decise perciò di chiamare l'avvocato.

punto di abbreviazione *punto di fine frase*

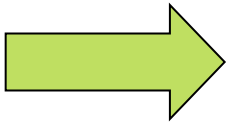
Sentence Splitting

- Feature *utilizzate*:
  - **Feature Locali**:
    - Forma, Lunghezza del token, Presenza di punteggiatura all'interno del token (es Acronimi C.N.R.), Posizione della parola all'interno della frase, Presenza della parola all'interno di una lista di parole che noi consideriamo abbreviazioni ad alta probabilità
  - **Feature Contestuali**:
    - Token precedente, Token successivo, Caratteristiche tipografiche della parola successiva (es. inizia con una maiuscola)
  - Di solito non si usano **feature Globali**

Tokenizzazione

- Mediante il processo di “tokenizzazione” il testo viene segmentato in unità ortografiche.
- Compito di solito affrontato con approcci a regole (espressioni regolari).

Il danno non poteva essere sottovalutato ...



1	Il
2	danno
3	non
4	Poteva
5	essere
6	sottovalutato

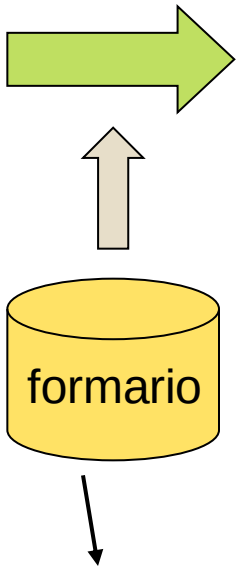
Criticità

- riconoscimento di “multiword” (ad hoc, ex aequo, ecc.)
- gestione di unità non lessicali (date, elementi numerici, emoticons, ecc.)

Analisi morfologica

- Alle unità ortografiche (token) sono associate tutte le possibili letture morfologiche utilizzando un dizionario delle forme (o *formario*)

id	forma
1	Il
2	danno
3	non
4	poteva
5	essere
6	sottovalutato



id	forma	lemma	pos	tratti
1	Il	il	RD	MS
2	danno	danno;dare	S;V	MS;P3IP
3	non	non	B	NULL
4	poteva	potere	V	S3II
5	essere	essere	V	F
6	sottovalutato	sottovalutare	V	MSPR

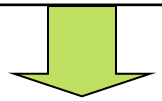
costituito da milioni di forme

schema di rappresentazione tabellare ”CoNLL”

Analisi morfo-sintattica (POS-tagging)

- Il PoS Tagging è il processo di disambiguazione morfologica.

id	forma	lemma	pos	tratti
1	Il	il	RD	MS
2	danno	danno;dare	S;V	MS;P3IP
3	non	non	B	NULL



id	forma	lemma	pos	tratti
1	Il	il	RD	MS
2	danno	danno	S	MS
3	non	non	B	NULL

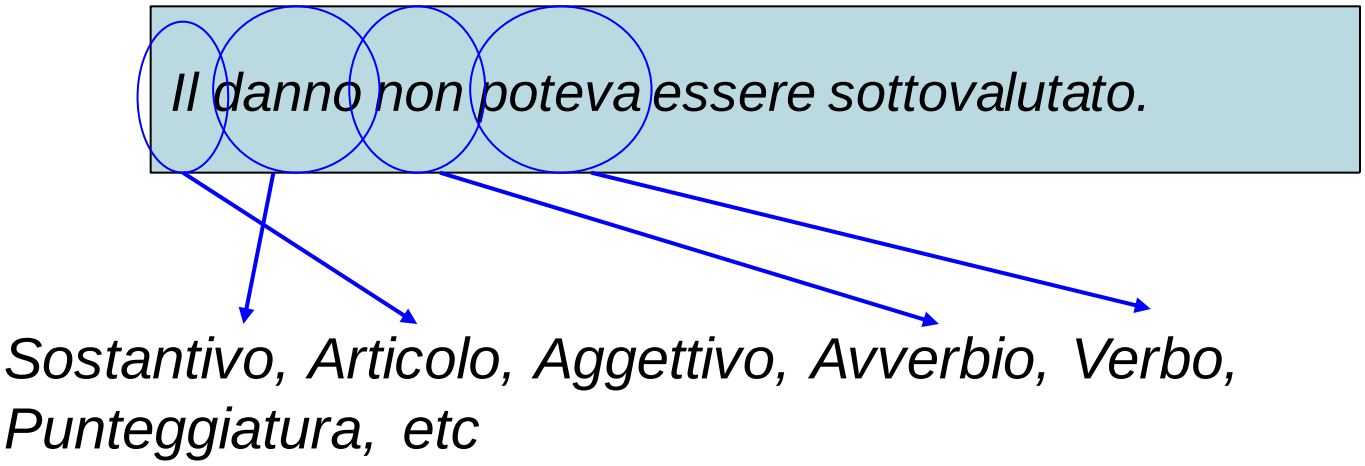
Criticità

- disambiguazione tra sostantivo-aggettivo (es: Il paziente inglese), aggettivo-participio passato (es: Disegno colorato dal paziente inglese)

Part-of-Speech Tagging

Nel caso dell'analisi morfo-sintattica (POS-tagging) il compito dell'analisi grammaticale diventa quello di assegnare ad ogni token della frase la giusta classe grammaticale:

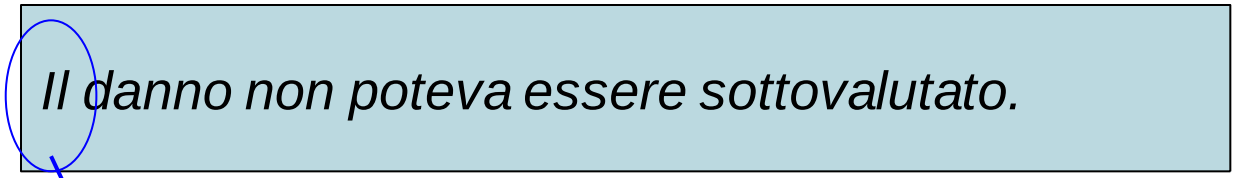
Sostantivo, Aggettivo, Avverbio, Verbo, Punteggiatura, Articolo, etc



27

Part-of-Speech Tagging

Il sistema si complica quando dobbiamo determinare anche I tratti morfologici (genere, numero, tempo, modo, etc.) per ogni parola. Tali tratti generano un numero maggiore di classi:



Articolo-Maschile-Singolare, Articolo-Femminile-Singolare, Articolo-Maschile-Plurale, Articolo-Femminilie-Plurale, etc..

28

Part-of-Speech Tagging

- Feature *utilizzate*:
  - **Feature Locali**:
    - Forma, Lunghezza del token, Presenza di punteggiatura all'interno del token (es Acronimi C.N.R.), Prefisso, Suffisso, Caratteristiche tipografiche del token
  - **Feature Contestuali**:
    - Token precedente, Token successivo, Risultato dell'analisi del token precedente, Possibili classi grammaticali della parola successiva (estratti dal livello di analisi morfologica ambigua)
  - Di solito non si usano **feature Globali**

Part-of-Speech Tagging: TagSet

- Tagset utilizzato in EVALITA 2009: definito all'interno di un progetto congiunto tra Dipartimento di Informatica dell'Università di Pisa e l'Istituto di Linguistica Computazionale
- Tre livelli di POS tags: *coarse-grained*, *fine-grained* e *morphed tags*
- *coarse-grain*, 14 categorie:

Tag	Descrizione
A	*Aggettivo*
B	*Avverbio*
C	*Congiunzione*
D	*Determinante*
E	*Preposizione*
F	*Punteggiatura*
I	*Interiezione*
N	*Numerale*
P	*Pronome*
R	*Articolo*
S	*Nome*
T	*Pre-Determinante*
V	*Verbo*
X	*Classe Residua*

Part-of-Speech Tagging: TagSet

- *fine-grained,* 36 categorie:

Tag	Descrizione
A	*Aggettivo*
AP	*Aggettivo Possessivo*
B	*Avverbio*
BN	*Avverbio di negazione*
…	…
S	*Nome Comune*
SA	*Nome Abbreviato*
SP	*Nome Proprio*
…	…
Vip	*Verbo Principale Indicativo Presente*
Vii	*Verbo Principale Indicativo Imperfetto*
…	…

Part-of-Speech Tagging: TagSet

- *morphed tags*: 328 categorie: *fine-grained* + genere, numero, persona, modo, tempo, presenza del clitico.

Tag	Descrizione
Ams	*Aggettivo Maschile Singolare*
Afs	*Aggettivo Femminile Singolare*
Amp	*Aggettivo Maschile Plurale*
Afp	*Aggettivo Femminile Plurale*
B	*Avverbio*
…	…
Sms	*Nome Comune Maschile Singolare*
Sfs	*Nome Comune Femminile Singolare*
…	…
SP	*Nome Proprio*
…	…
VAip3s	*Verbo ausiliare indicativo presente terza persona singolare*
…	…

Part-of-Speech Tagging

valutazione

- L'accuratezza del sistema è lo stato dell'arte per l'italiano (Evalita-2009 PoSTagging Task):

global data		unknown tokens	
accuracy	**error rate**	**accuracy**	**error rate**
96.34%	3.66%	91.07%	8.93%

- Errori più frequenti:

Our result -> Correct	% Error rate
ADJ ->NN	9.8%
NN -> ADJ	9.3%
V_PP -> ADJ	8.1%

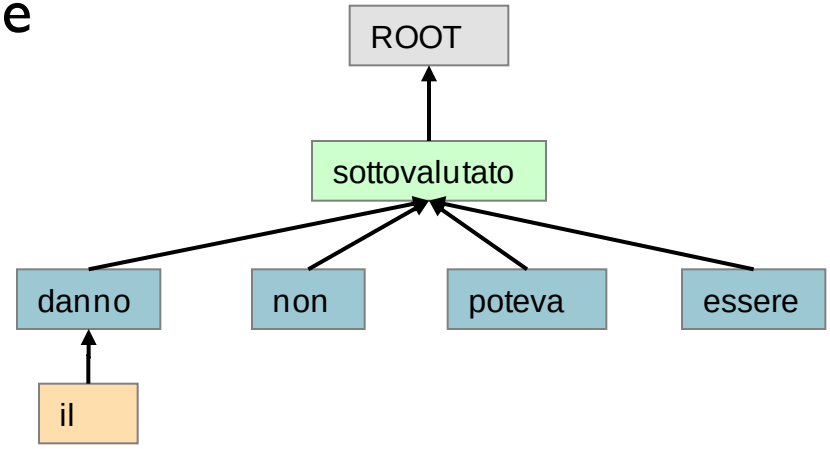
Maximum Entropy for Italian Pos Tagging. Dell'Orletta F., Federico M., Lenci A., Montemagni S., Pirrelli V. In: *Proceedings of Workshop Evalita 2007*. *Intelligenza Aritificiale*, 4(2), 2007.

Embedded System for Pos Tagging. Dell'Orletta F. In: *Proceedings of Evalita 2009.*

Parsing Sintattico a Dipendenze

In questa compito di analisi vengono individuate le relazioni sintattiche tra i token della frase

id	forma	lemma	pos	tratti
1	Il	il	RD	MS
2	danno	danno	S	MS
3	non	Non	B	NULL
4	poteva	Potere	V	S3II
5	essere	essere	V	F
6	sottovalutato	sottovalutare	V	MSPR



id	forma	lemma	pos	tratti	head	dep
1	Il	il	RD	MS	2	DET
2	danno	danno	S	MS	6	SUBJ_PASS
3	non	non	B	NULL	6	NEG
4	poteva	potere	V	S3II	6	MODAL
5	essere	essere	V	F	6	AUX
6	sottovalutato	sottovalutare	V	MSPR	0	ROOT

Dependency Parsing come Problema di Classificazione

- Esistono diversi metodi per trasformare un compito di analisi sintattica in un compito di classificazione, sicuramente uno dei metodi più famosi è quello proposto da Yamada e Matsumoto nel 2003, chiamato ***Shift/Reduce parser*** (o ***transition-based parser***) parser
- Il compito di analizzare sintatticamente una frase diventa il compito di predire l'azione che il parser deve fare per costruire l'albero sintattico della frase
- Ad ogni passo dell'analisi il parser usa un classificatore addestrato su una **TreeBank** (o training corpus) allo scopo di predire quale azione deve compiere dato l'insieme delle feature (*locali+contestuali*) estratte in quel determinato momento

Dependency Parsing come Problema di Classificazione

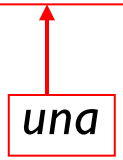
- Il parser costruisce l'albero a dipendenza analizzando la frase da sinistra verso destra e compiendo tre *azioni*: ***Shift***, ***Right*** e ***Left:***

  - ***Shift***: non c'è nessuna relazione tra le due parole target analizzate, l'analisi si muove verso ***destra:***

Io vidi una donna con gli occhiali --> io vidi una donna con gli occhiali

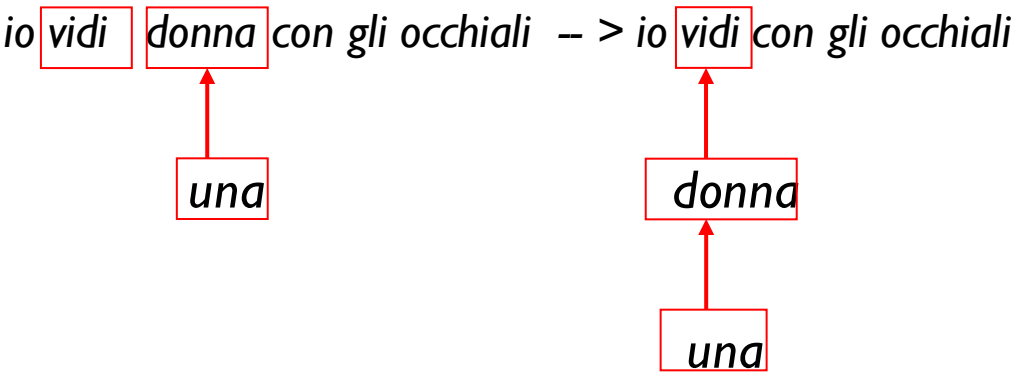
  - ***Right***: esiste una relazione tra le due parole, il nodo sinistro è considerato ***dipendente*** del nodo ***testa*** a destra

Io vidi una donna con gli occhiali --> io vidi donna con gli occhiali



Dependency Parsing come Problema di Classificazione

- ***Left***: esiste una relazione tra le due parole, il nodo sinistro è considerato ***testa*** del nodo ***dipendente*** a desta



- L'algoritmo va avanti fino a quando non è stato completato l'albero: sono stati creati tutti i link sintattici

Dependency Parsing come Problema di Classificazione

- A questo punto abbiamo ottenuto un albero sintattico ***non marcato*** (gli archi non sono marcati con le relazioni di dipendenza: soggetto, oggetto, complemento di tempo, etc).
- Come possiamo fare per ottenere un albero marcato? Esistono almeno due modi:

  • attraverso un secondo passo di analisi nel quale si classifica ogni arco con la classe sintattica più probabile (problema di classificazione)

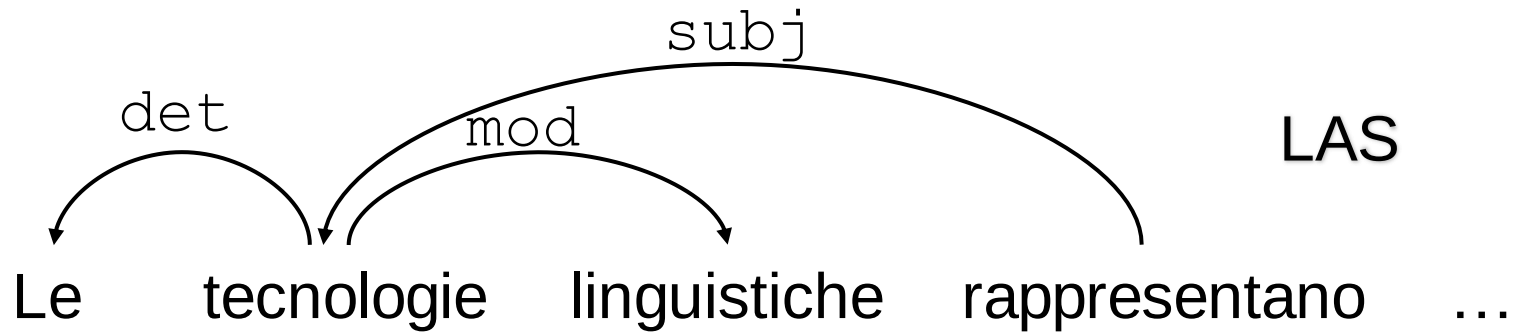
  • semplicemente aumentando il numero delle azioni del parser, non più solo Shift, Right e Left, ma:
  Shift, *Right_soggetto*, *Right_oggetto*, *Right_comp_di_tempo*, …, *Left_soggetto*, *Left_oggetto*, *Left_comp_di_tempo*, …..

Dependency Parsing come Problema di Classificazione

- Quindi il compito di analisi sintattica di una frase diventa un compito di classificazione che può essere diviso in tre fasi:
  - ***estrazione delle feature*** (locali e contestuali) rispetto alle due parole analizzate
  - **stima dell'azione da eseguire** attraverso l'algoritmo di apprendimento supervisionato (dato un modello di feature pesate)
  - **esecuzione dell'azione** e creazione dell'albero

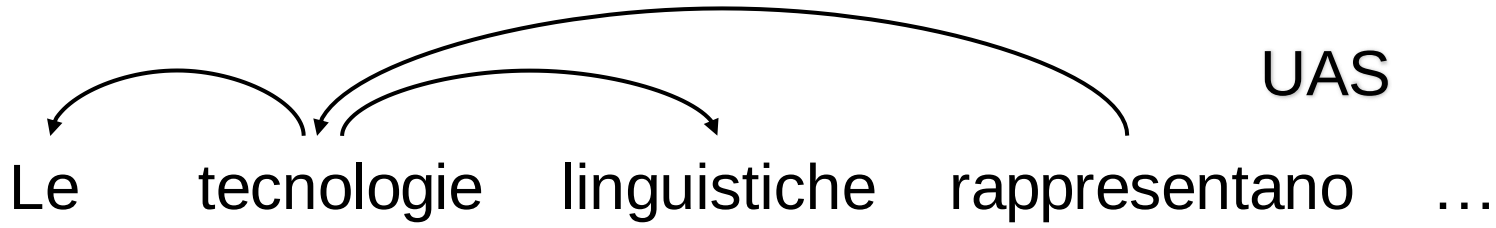
Dependency Parsing: valutazione

- Metrica ufficiale di valutazione:
  - **LAS** (Labeled Accuracy Score): *percentuale di dipendenze identificate ed etichettate correttamente*
- Altre metriche di valutazione:
  - **UAS** (Unlabeled Accuracy Score): *percentuale di dipendenze identificate correttamente*
  - **LA** (Label Accuracy Score): *percentuale di dipendenze etichettate correttamente*



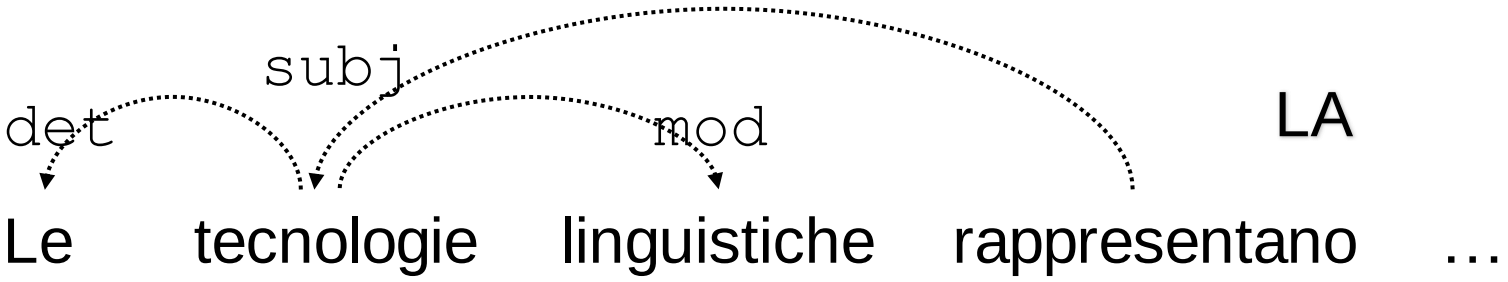
Dependency Parsing: valutazione

- Metrica ufficiale di valutazione:
  - **LAS** (Labeled Accuracy Score): *percentuale di dipendenze identificate ed etichettate correttamente*
- Altre metriche di valutazione:
  - **UAS** (Unlabeled Accuracy Score): *percentuale di dipendenze identificate correttamente*
  - **LA** (Label Accuracy Score): *percentuale di dipendenze etichettate correttamente*



Dependency Parsing: valutazione

- Metrica ufficiale di valutazione:
  - **LAS** (Labeled Accuracy Score): *percentuale di dipendenze identificate ed etichettate correttamente*
- Altre metriche di valutazione:
  - **UAS** (Unlabeled Accuracy Score): *percentuale di dipendenze identificate correttamente*
  - **LA** (Label Accuracy Score): *percentuale di dipendenze etichettate correttamente*



Parsing a Dipendenze:

valutazione

- Stato dell'arte per l'Italiano:

ISST-TANL	
LAS	**UAS**
83.38%	87.71%

- Errori più frequenti:

Dipendenze	Recall	Precision	Error-rate
Comp_temp	29.41%	66.67%	0.3%
Comp_loc	40.24%	63.46%	1.6%
Con	59.70%	61.86%	3.1%
Arg	61.80%	66.27%	2.8%
Subj_pass	56.52%	76.47%	0.3%
Subj	82.86%	80.56%	4.8%
Obj	91.93%	80.00%	4.4%

***Reverse Revision and Linear Tree Combination for Dependency Parsing*".** Attardi G., Dell'Orletta F. " In: NAACL-HLT 2009 , Boulder, Colorado.

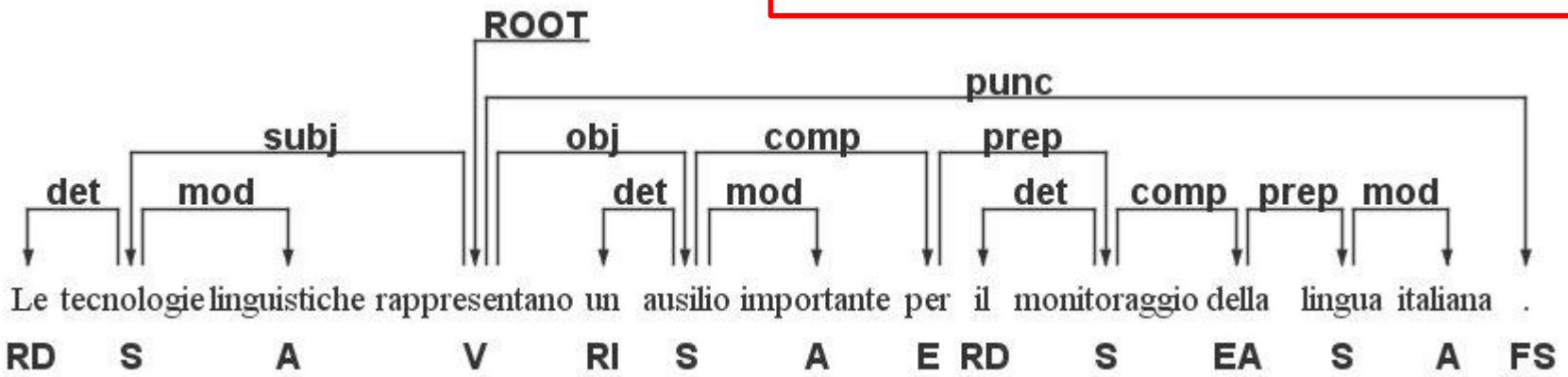
Le tecnologie linguistiche

Annotazione sintattica a dipendenze

Conll-2007: 81.3% LAS

Evalita 2009: 83.38% LAS

Stato dell'arte per l'italiano

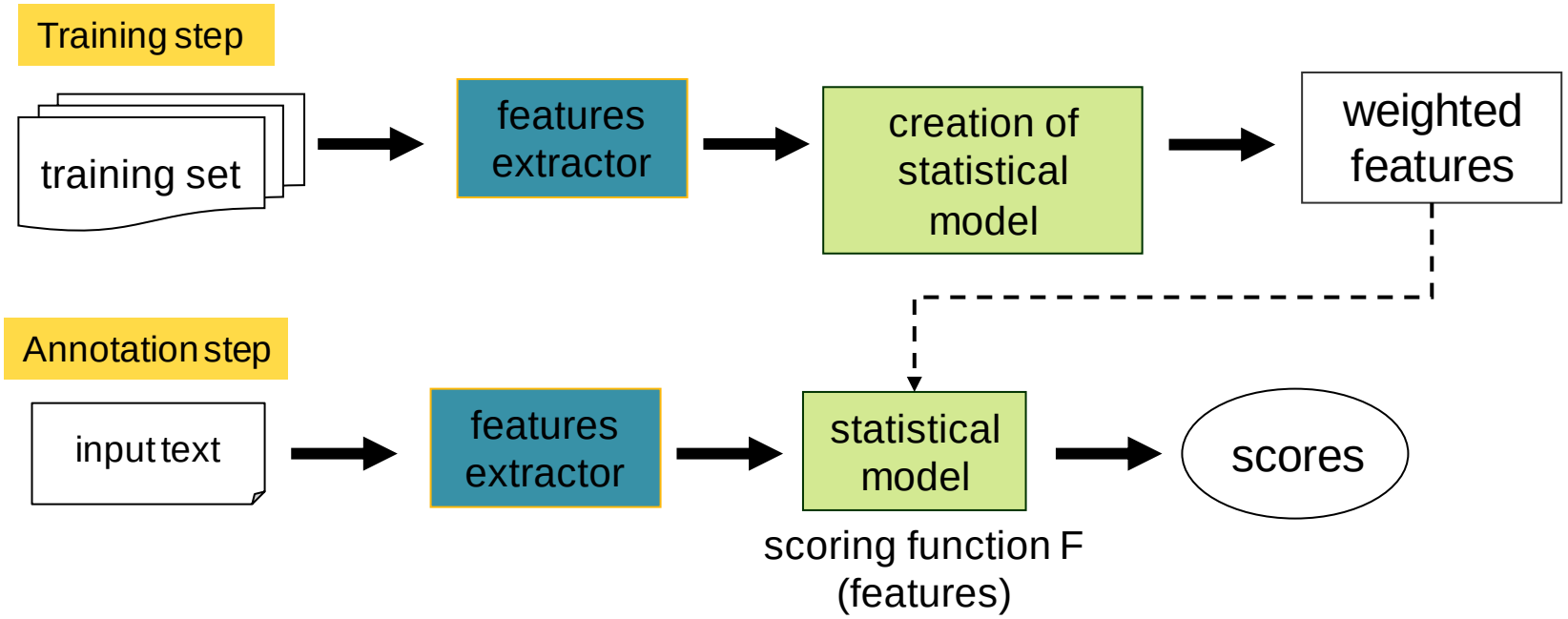


Annotazione morfo-sintattica

Evalita 2009: accuratezza = 96,34%

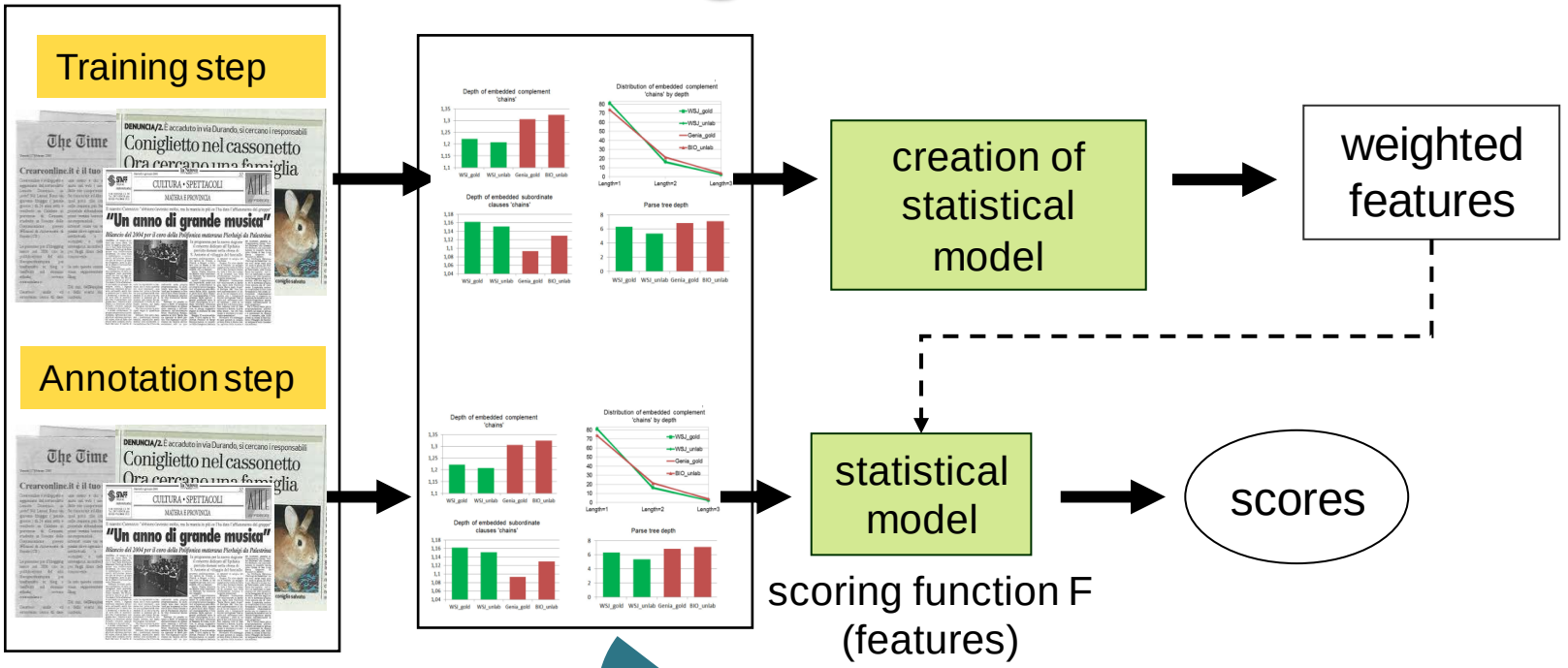
Stato dell'arte per l'italiano

Annotazione linguistica stocastica



- 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 viene poi applicato a esempi sconosciuti per assegnare loro la classe più probabile, dato il modello e l'insieme di tratti pertinenti.
- **scoring function**: usa sia le "weighted features" sia le "extracted features" per identificare la classificazione più probabile

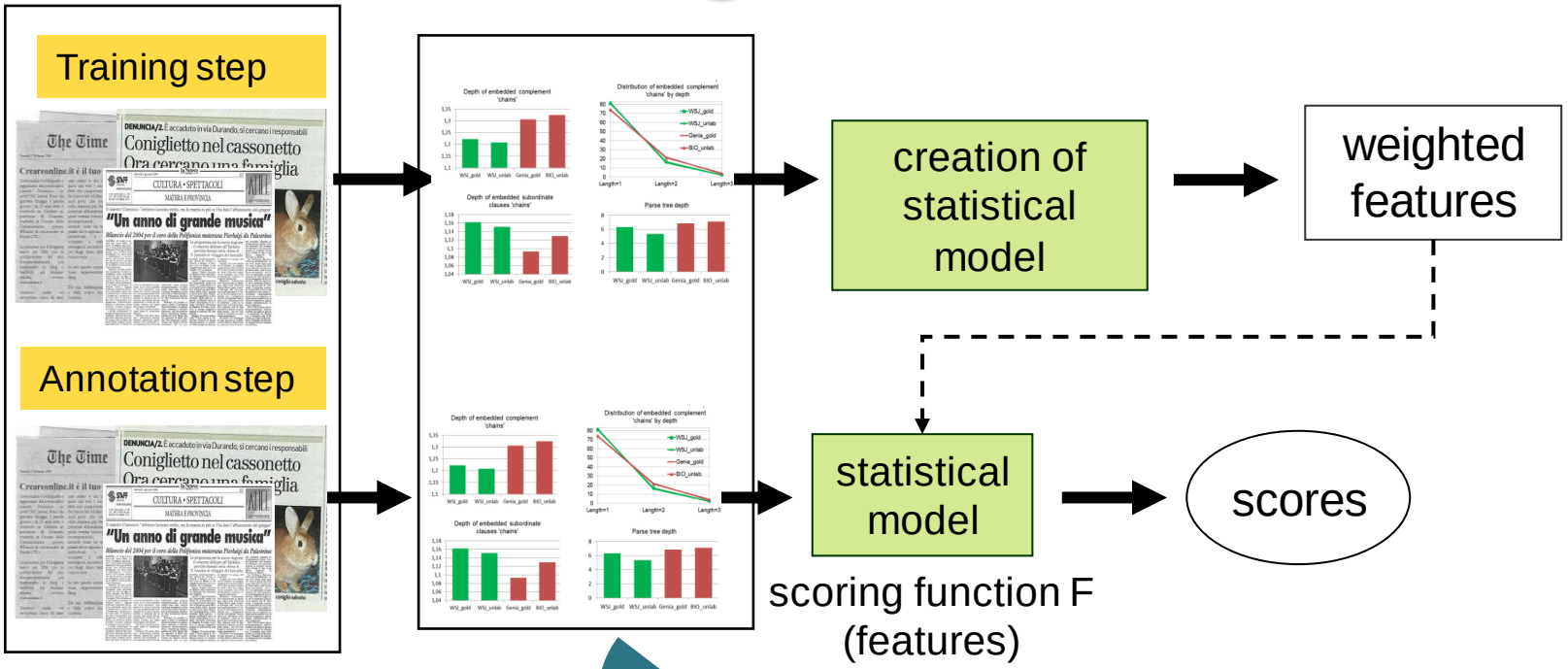
Annotazione linguistica stocastica



- Il campione di addestramento e il testo sconosciuto appartengono allo stesso dominio
- Gli strumenti di annotazione stocastica sono tipicamente addestrati su corpora giornalistici

- Il campione di addestramento e il testo sconosciuto condividono la stessa distribuzione di tratti contestuali e linguistici
- Sono tratti tipicamente rappresentativi del linguaggio giornalistico

Annotazione linguistica stocastica

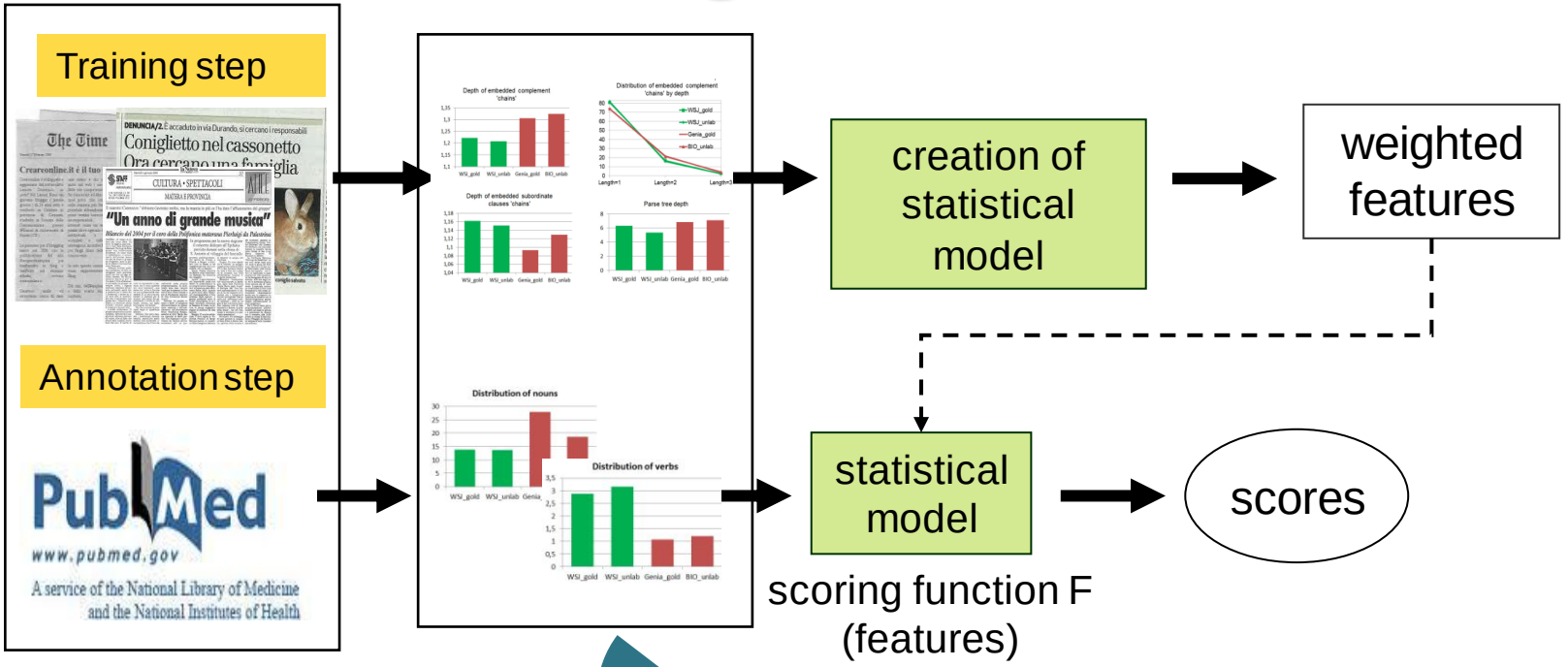


- Il campione di addestramento e il testo sconosciuto appartengono allo stesso dominio
- Gli strumenti di annotazione stocastica sono tipicamente addestrati su corpora giornalistici

- Buon livello di accuratezza
- Es.: DeSR parser addestrato e testato sulla PennTreebank

Test corpus	LAS	UAS
PennTreebank	86.09%	87.29%

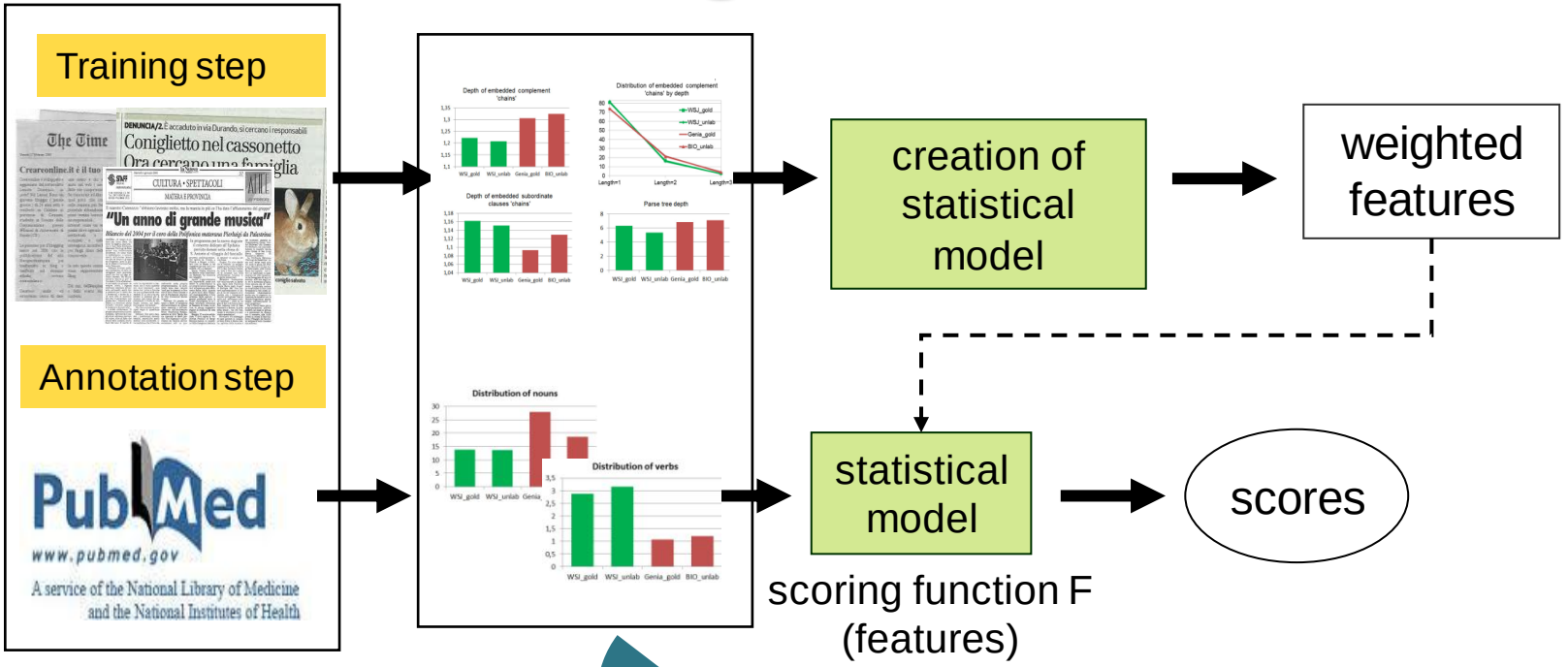
Annotazione linguistica stocastica



- Il campione di addestramento e il testo sconosciuto appartengono a due domini diversi
- Es.: addestramento su corpora giornalistici e annotazione di articoli biomedici *(inglese)*

- Diversa distribuzione di tratti contestuali e linguistici
- Es.: addestramento rispetto a tratti del linguaggio giornalistico e annotazione di testi rappresentativi del linguaggio biomedico

Annotazione linguistica stocastica



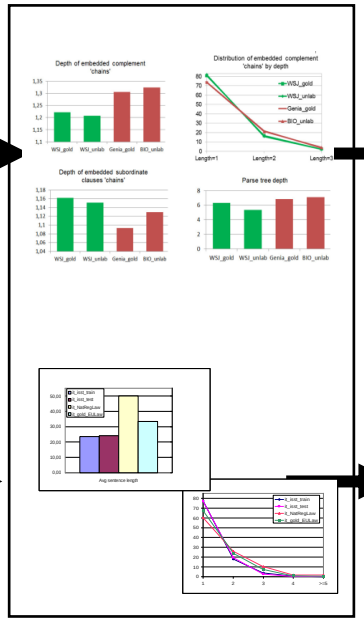
- Diminuzione di accuratezza

LAS: -7.5%

UAS: -6% (CHEM), -7% (BIO e GENIA)

Test corpus	LAS	UAS
PennTreebank	86.09%	87.29%
CHEM	78.50%	81.10%
BIO	78.65%	79.97%
GENIA	n/a	80.25%

Annotazione linguistica stocastica



creation of statistical model

weighted features

statistical model

scores

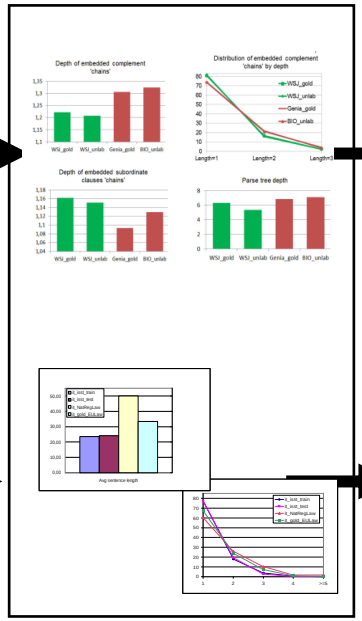
scoring function F (features)

- Il campione di addestramento e il testo sconosciuto appartengono a due domini diversi
- Es.: addestramento su corpora giornalistici e annotazione di testi giuridici *(italiano)*

- Diminuzione di accuratezza
- Es.: DeSR and MST parser addestrati sulla ISST-TANL Treebank e testati su un corpus di testi giuridici

Test corpus	DeSR:LAS	MST:LAS
ISST-TANL	82.09%	75.85%
Testi giuridici	75.85%	74.62%

Annotazione linguistica stocastica



creation of statistical model

weighted features

statistical model

scores

scoring function F (features)

- Diminuzione di accuratezza

Es.: DeSR and MST parser addestrati sulla ISST-TANL Treebank e testati su un corpus di testi giuridici

LAS: -6.24% (DeSR)

-5.57% (MST)

Test corpus	DeSR:LAS	MST:LAS
ISST-TANL	82.09%	75.85%
Testi giuridici	75.85%	74.62%

Domain adaptation: il problema

- Gli strumenti di annotazione linguistica stocastica hanno una notevole diminuzione del livello di accuratezza quando sono testati su testi con caratteristiche diverse da quelle del training
- Scenario d'uso reale:
  - annotazione di testi profondamente diversi dal training
  - 

Il problema del **Domain Adaptation**: la necessità di adattare gli strumenti sviluppati sulla base di un dominio d'origine all'analisi di un nuovo dominio target

•Barbara Plank: http://cst.dk/bplank/proefschrift/thesis-bplank.pdf

•David McClosky: http://nlp.stanford.edu/~mcclosky/papers/dmcc-thesis-2010.pdf

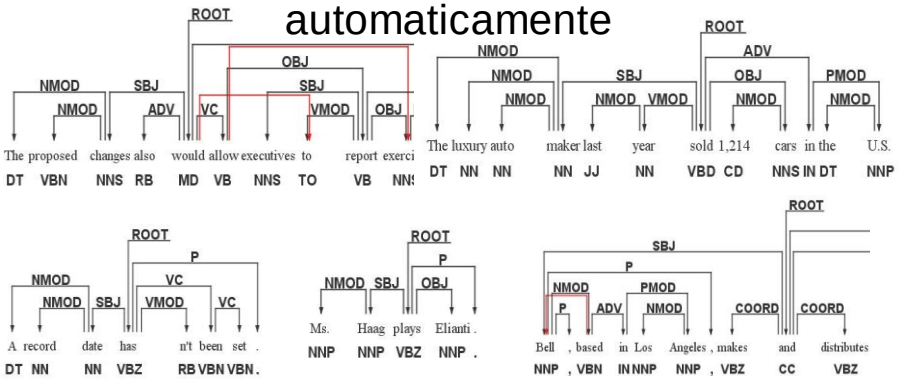
Self-training per domain adaptation

- Algoritmo di self-training basato su ULISSE: algoritmo capace di selezionare da una grande quantità di testi annotati automaticamente le analisi corrette corrispondenti alle frasi più informative
  - ULISSE associa un punteggio di accuratezza ad ogni frase analizzata sintatticamente e crea un ranking delle frasi analizzate. *(Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni (2011), ULISSE: an Unsupervised Algorithm for Detecting Reliable Dependency Parses (CoNLL 2011))*
  - Le frasi analizzate vengono unite al training originario del dominio di partenza (giornalistico)
  - Testato sul dominio biomedico
    - *Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni (2013), "Unsupervised Linguistically-Driven Reliable Dependency Parses Detection and Self-Training for Adaptation to the Biomedical Domain", ACL - BioNLP*

ULISSE

(Unsupervised LInguistically-driven Selection of dEpendency parses)

Ampio corpus di frasi annotate automaticamente



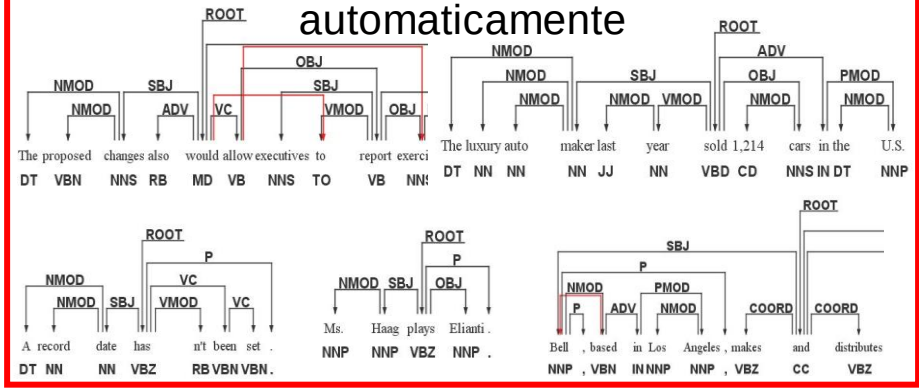
ULISSE crea un modello statistico utilizzando un insieme di **caratteristiche linguisticamente motivate** estratte dal corpus annotato automaticamente

Modello statistico

ULISSE

(*Unsupervised LInguistically-driven Selection of dEpendency parses*)

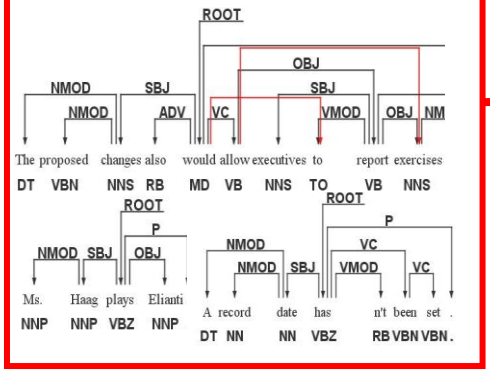
Ampio corpus di frasi annotate automaticamente



ULISSE crea un modello statistico utilizzando un insieme di **caratteristiche linguisticamente motivate** estratte dal corpus annotato automaticamente

Modello statistico

Frasi analizzate appartenenti allo stesso dominio del corpus

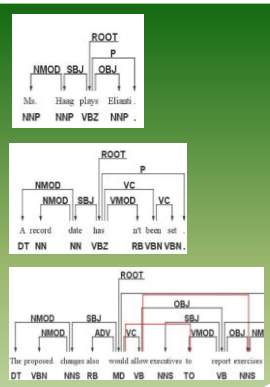


Modello statistico

ULISSE calcola un **punteggio di accuratezza** associato ad **ogni albero a dipendenza** per ogni frase analizzata

Ranking decrescente di analisi

(da corretti a scorretti)

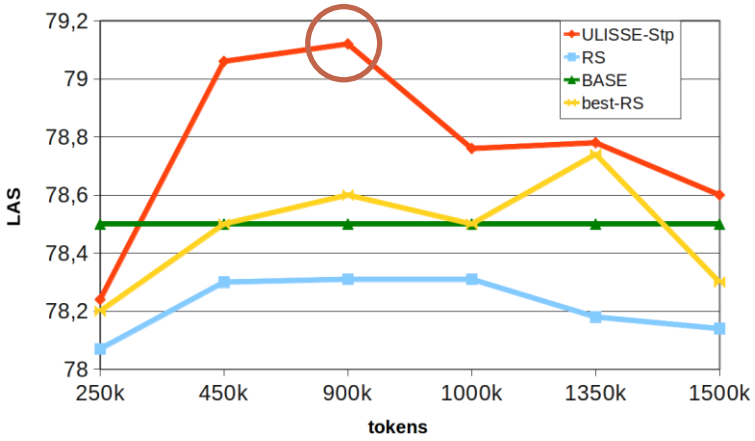


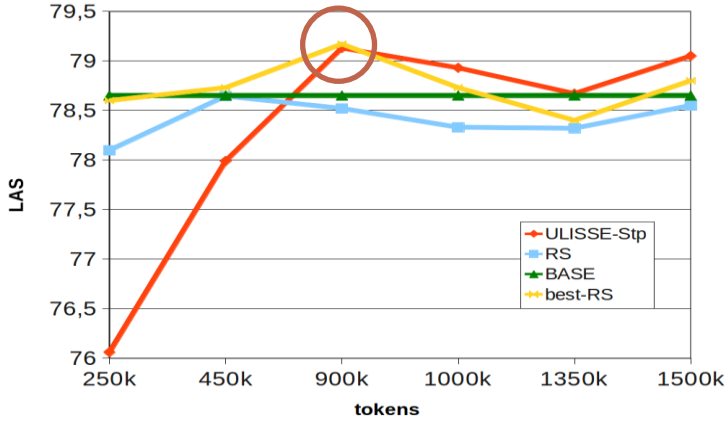
Risultati: CHEM e BIO

LAS per CHEM

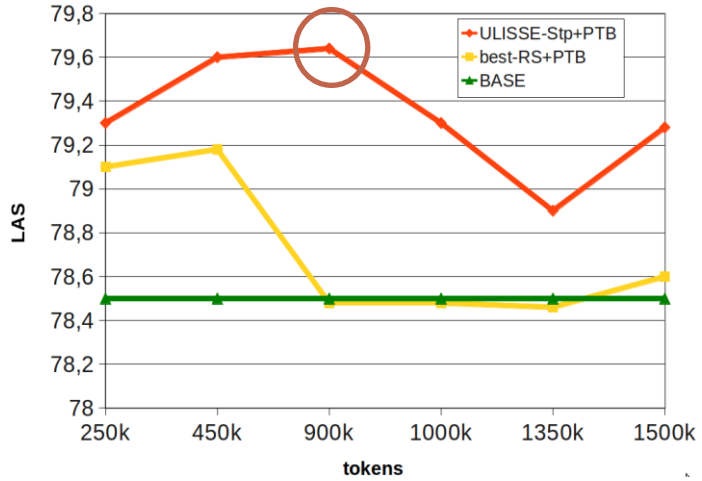
LAS per BIO

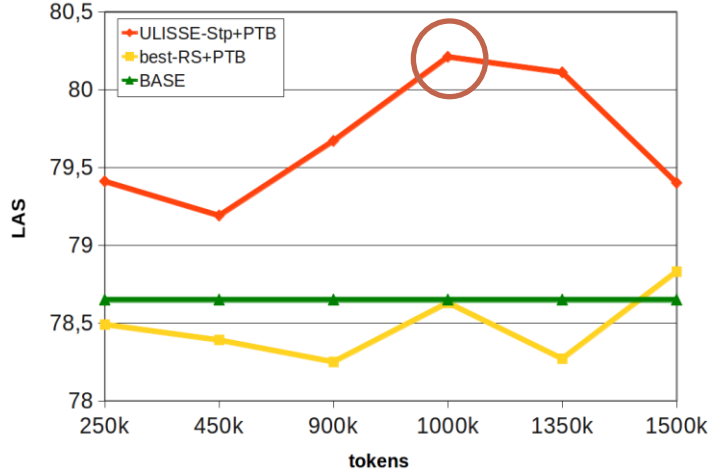
Senza PTB in addestramento





Con PTB in addestramento



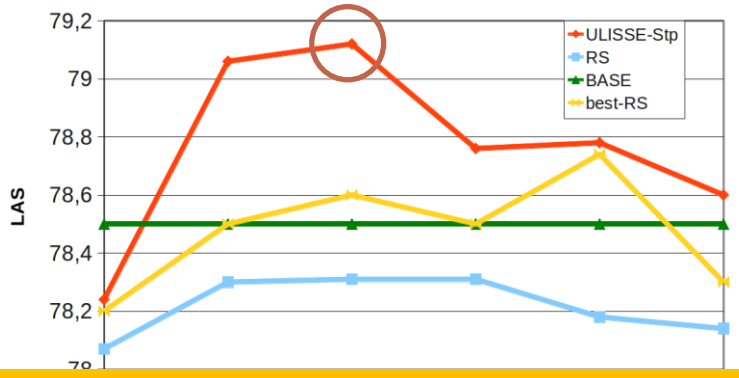


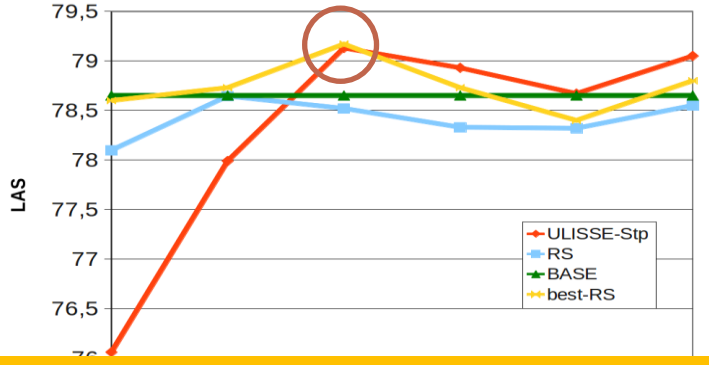
Risultati: CHEM e BIO

LAS per CHEM

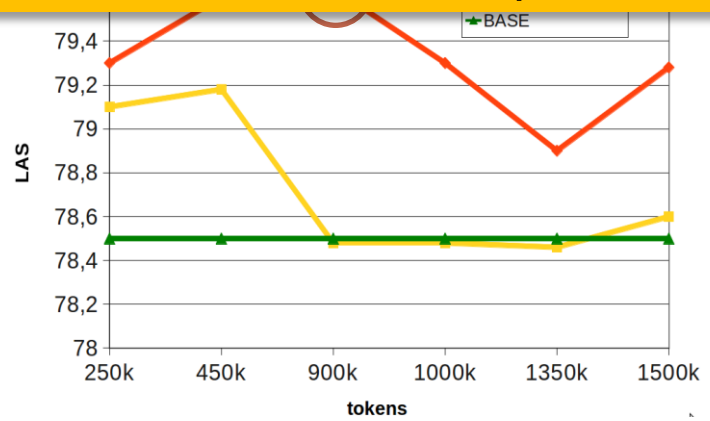
LAS per BIO

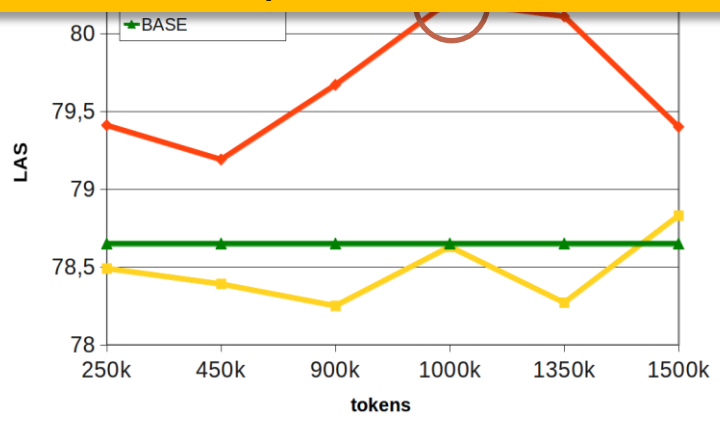
Senza PTB in addestramento





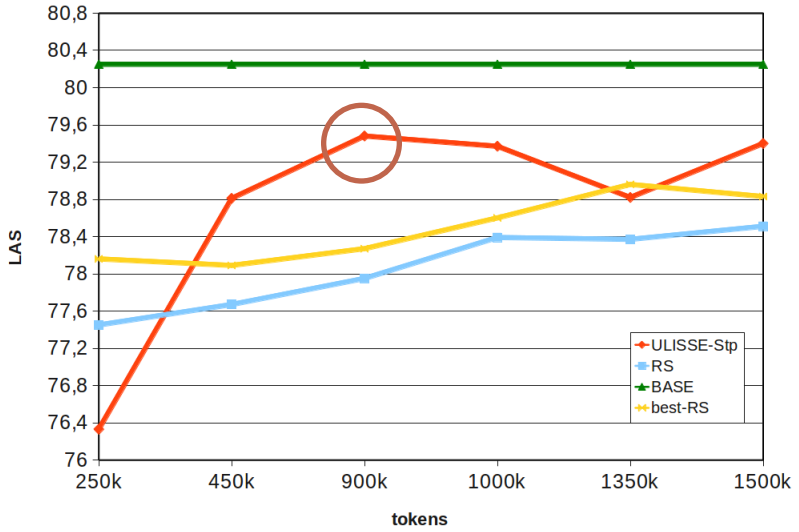
ULISSE–Stp usando in apprendimento solo dati analizzati automaticamente ottiene migliori risultati che il modello BASE (addestrato solo su PTB)



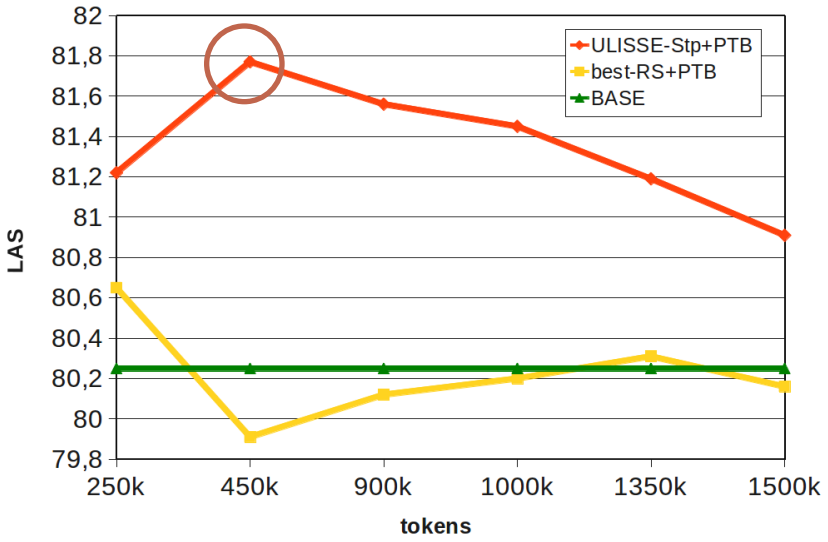


Risultati: Genia

- UAS per GENIA senza PTB in addestramento

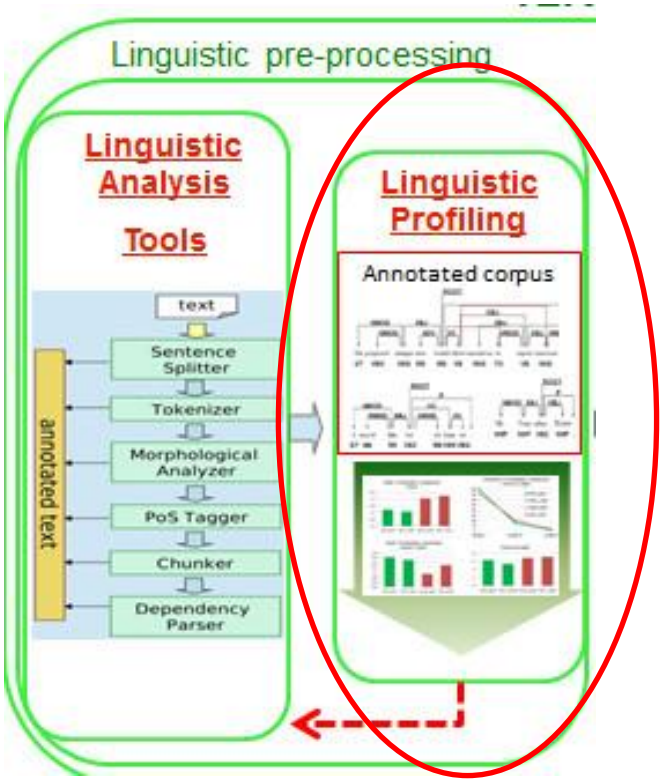


- UAS per GENIA con PTB in addestramento



In tutti i casi, le performance di ULISSE iniziano a decrescere quando un insieme troppo grande di frasi annotate automaticamente viene inserito in fase di addestramento

Annotazione del testo e monitoraggio linguistico in T2K



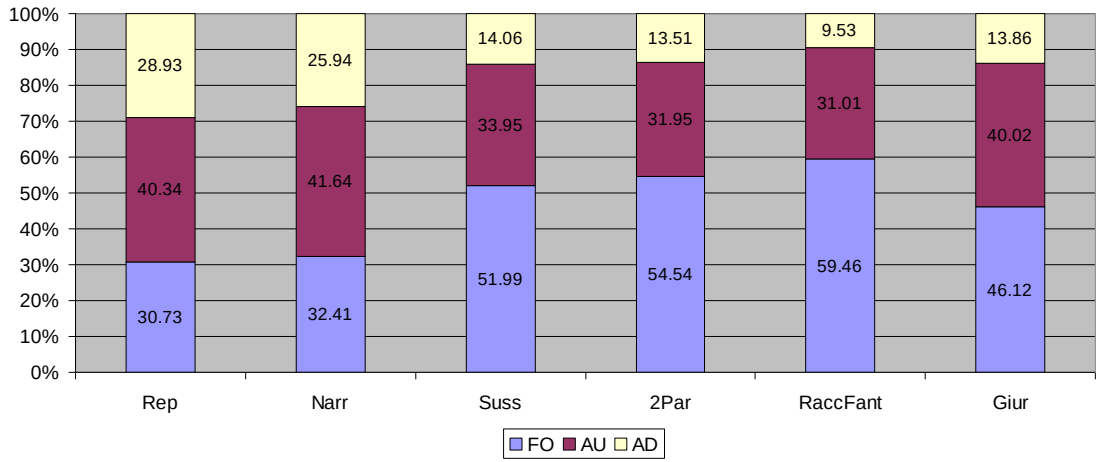
- The linguistically analyzed corpus is used by the **linguistic profiling module** to investigate the *form* of a text rather the *content*
- The distribution of a wide range of linguistic features (lexical, morpho-syntactic and syntactic) is aimed at
  - assessing the readability level (Dell'Orletta et al., 2011)
  - native language identification (Cimino et al., 2012)
  - determining the text genre (Dell'Orletta et al., 2013)
- Moreover, they can be used to refine the construction of the corpus
  - In terms of homogeneity and representativeness of a given domain

Selezione dei parametri di osservazione: analisi lessicale

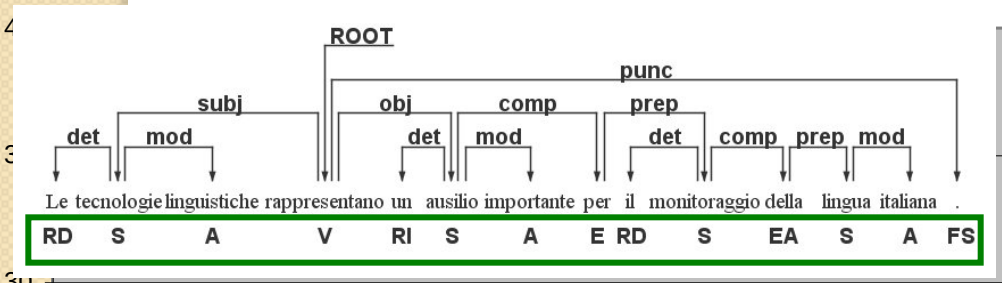
- Ripartizione del vocabolario appartenente al VdB rispetto ai repertori di uso FO, AU, AD

	Rep	Narr	Suss	2Par	RaccFant	Giur
Rapporto tipo/unità	**0.72**	**0.70**	0.68	**0.55**	0.18	0.38
Percentuale del vocabolario appartenente al VdB	**67.1**	71.76	73.57	74.58	**56.93**	35.60

- Ripartizione del vocabolario appartenente al VdB rispetto ai repertori di uso FO, AU, AD

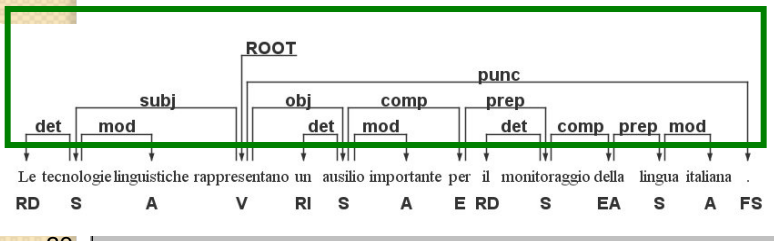


Parametri di osservazione: analisi morfo-sintattica

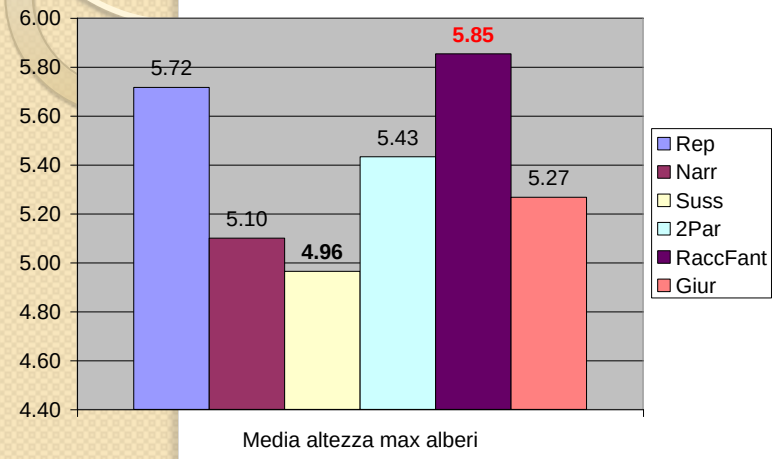


Distribuzione delle categorie morfo-sintattiche

Analisi sintattica: distribuzione dei tipi di dipendenza

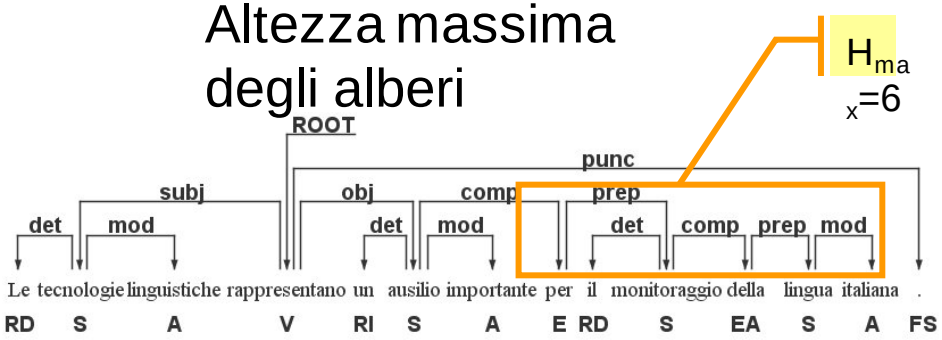


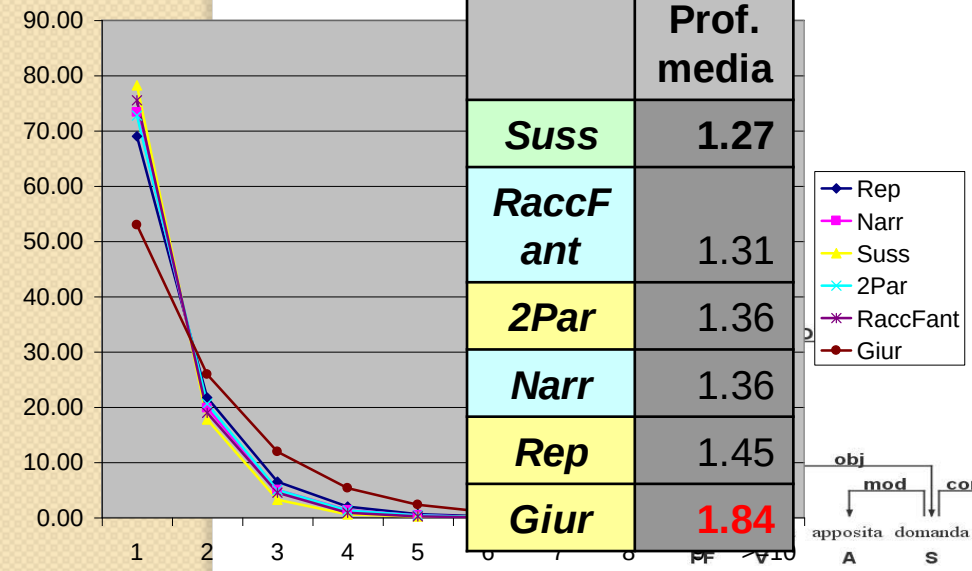
Analisi sintattica: parametri relativi ai livelli di incassamento



Media altezza max alberi

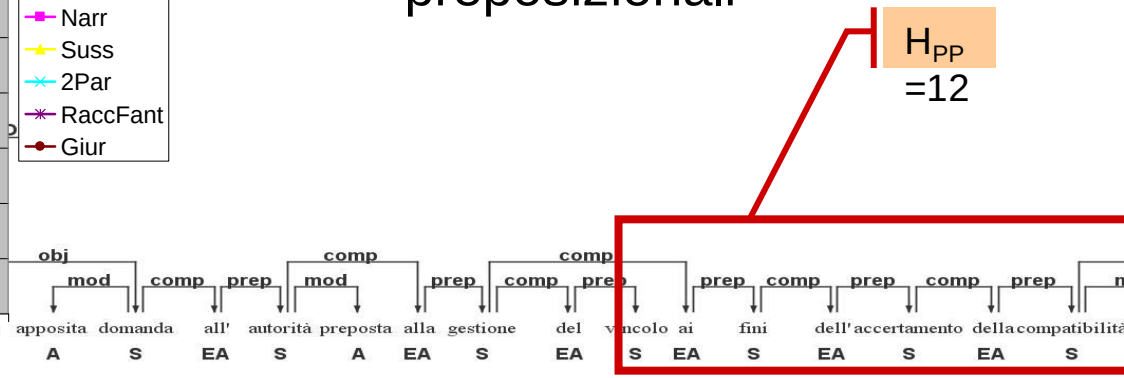
Altezza massima degli alberi





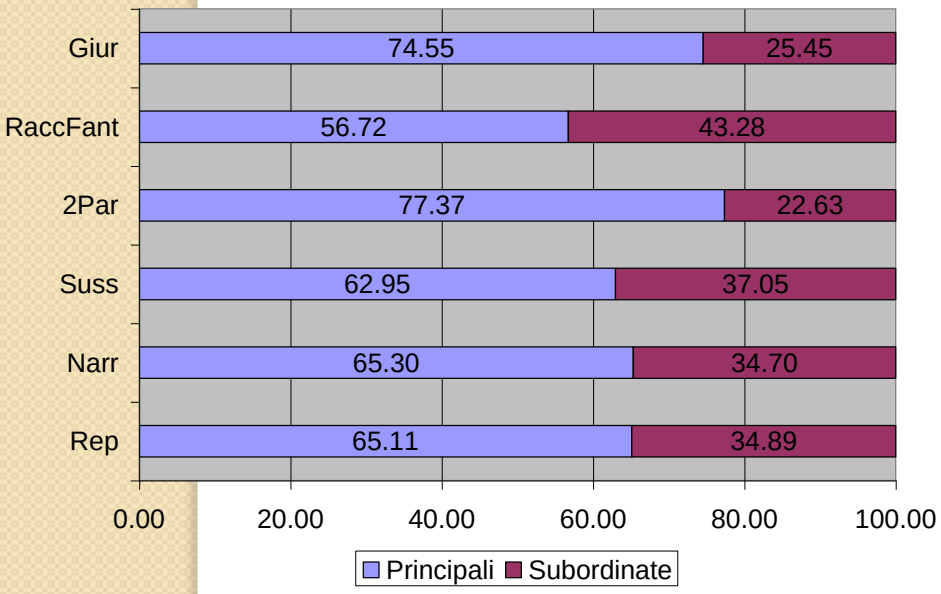
	Prof. media
Suss	1.27
RaccFant	1.31
2Par	1.36
Narr	1.36
Rep	1.45
Giur	1.84

Profondità delle “catene” di complementi preposizionali



Analisi sintattica: parametri relativi alla distribuzione delle teste verbali

	Rep	Narr	Suss	2Par	RaccFant	Giur
Media clausole/periodo	2.41	2.65	2.67	2.40	**3.37**	**1.64**



	Pre	Post
Rep	12.28	87.72
Narr	12.30	87.70
Suss	**13.03**	**86.97**
2Par	11.60	88.40
RaccFant	**5.58**	**94.42**
Giur	11.69	88.31

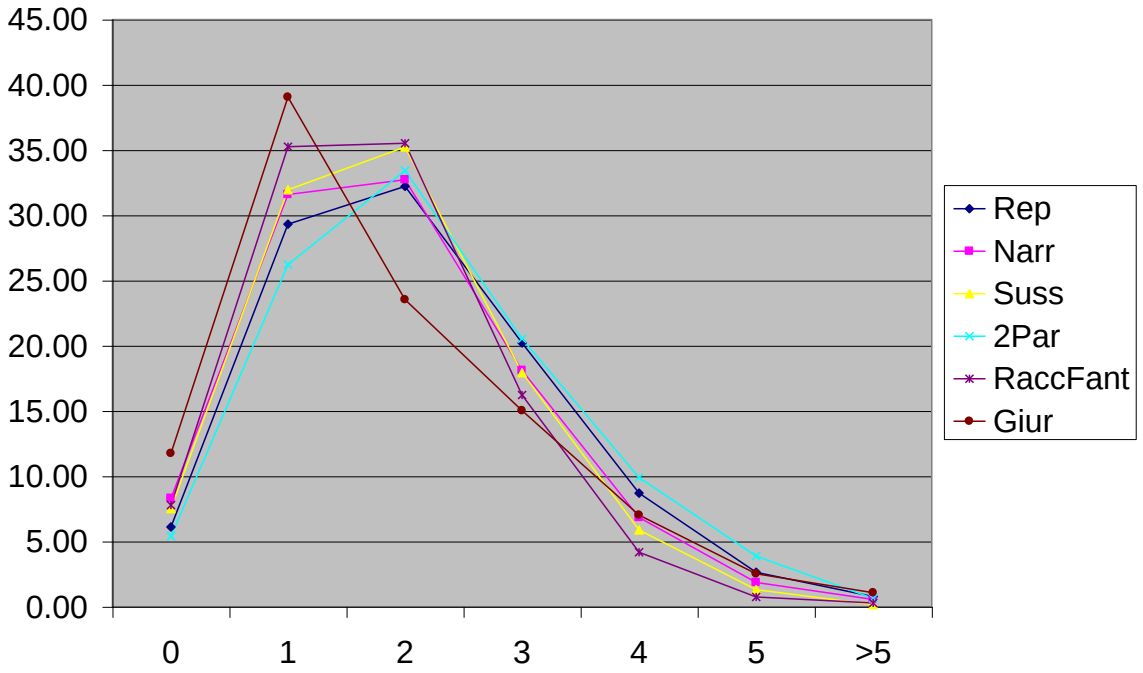
Ordine relativo delle subordinate rispetto alla principale

Analisi sintattica: parametri relativi alla distribuzione delle teste verbali

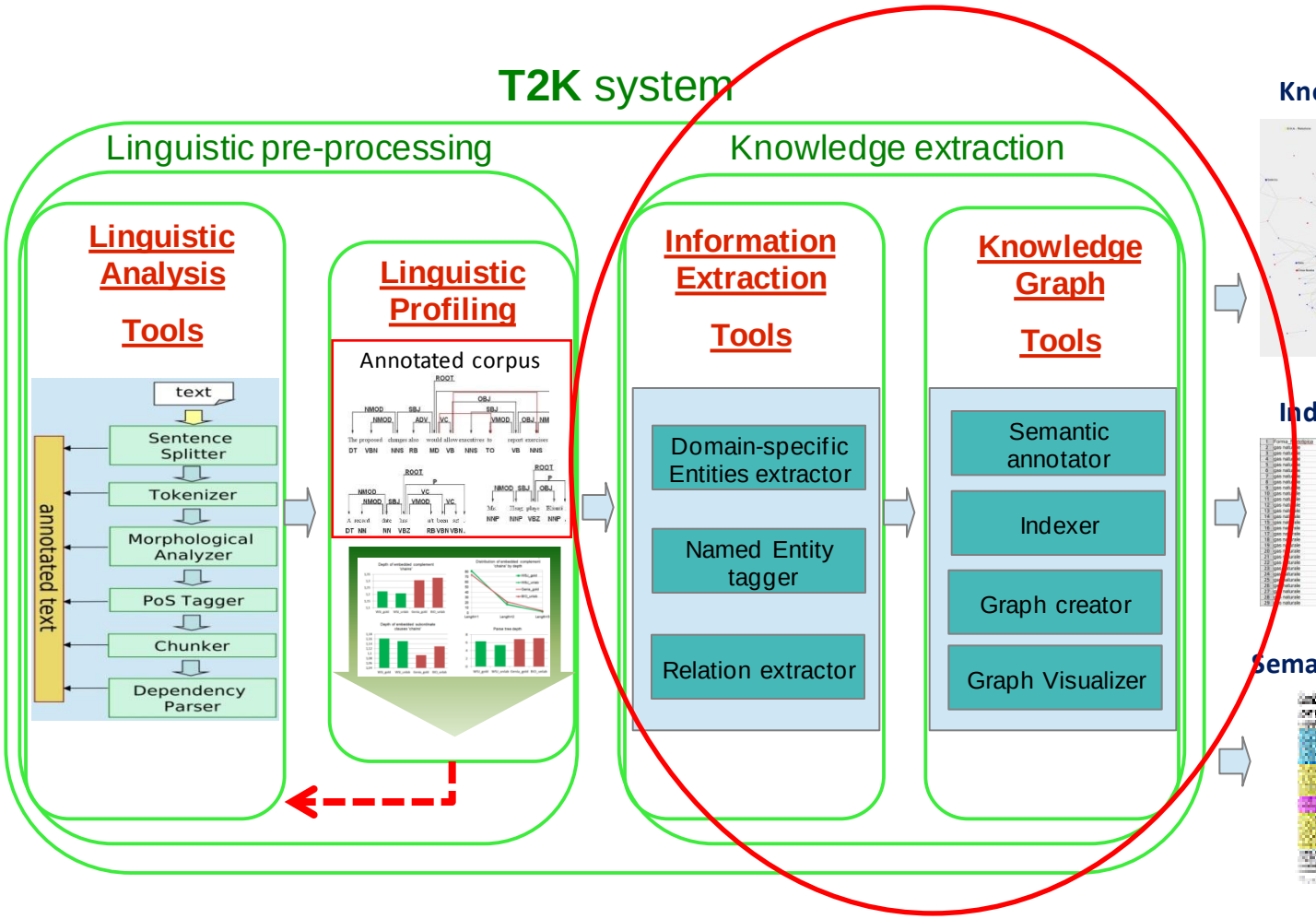
	Valenza media
Rep	2.07
Narr	1.92
Suss	1.87
2Par	2.18
RaccFant	**1.77**
Giur	1.79

Grado di “saturazione” delle valenze

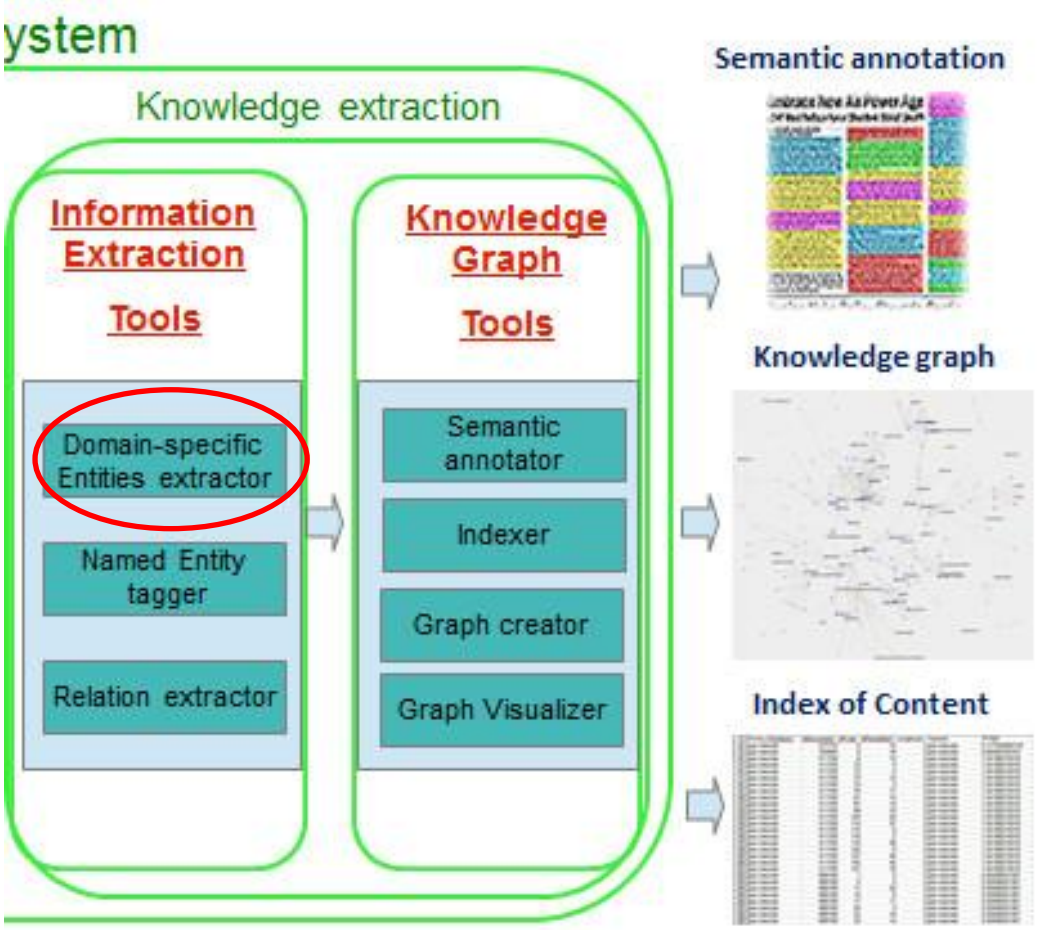
- “valenza” media verbale
- distribuzione dei verbi per “valenza”



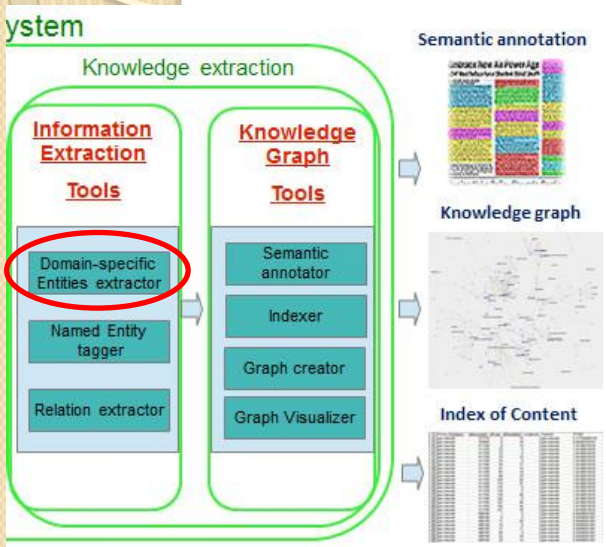
T2K: Estrazione di conoscenza di dominio



T2K: Estrazione di conoscenza di dominio



T2K: Estrazione di conoscenza di dominio

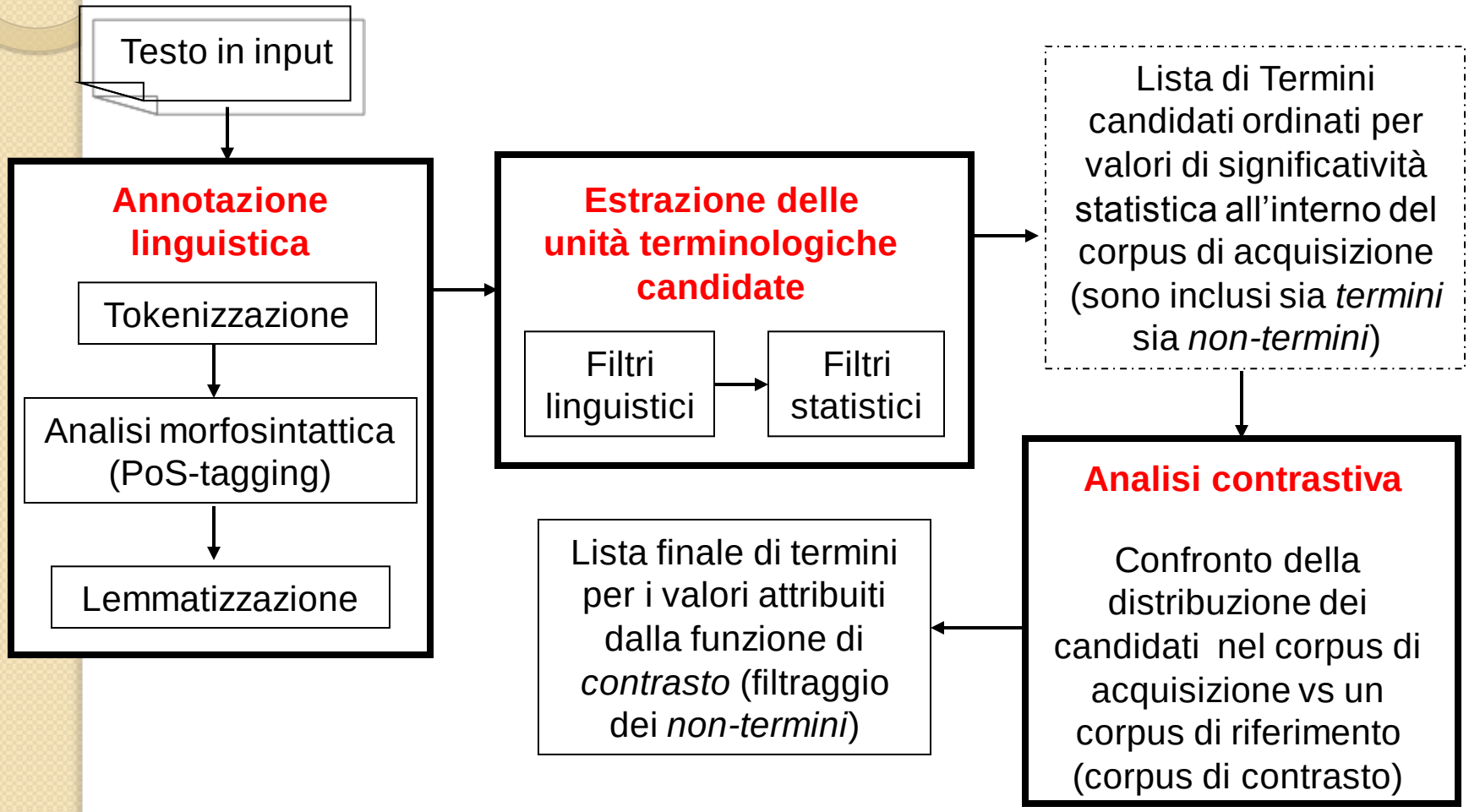


Input corpus: collezione di Direttive Europee in materia ambientale

Prototypical Form	Lemma of Term	Frequency
pile a combustibile	pila a combustibile	61
motore Diesel	motore Diesel	31
livelli trofici	livello trofico	27
nicchie ecologiche	nicchia ecologico	25
capacità portante	capacità portante	24
competizione interspecifica	competizione interspecifica	22
casa passiva	casa passivo	22
catena alimentare	catena alimentare	37
gas serra	gas serra	32
materia organica	materia organico	18
competizione intraspecifica	competizione intraspecifica	18
rapporto trofico	rapporto trofico	17
motore termico	motore termico	16
produzione primaria	produzione primario	16
anidride carbonica	anidride carbonico	46

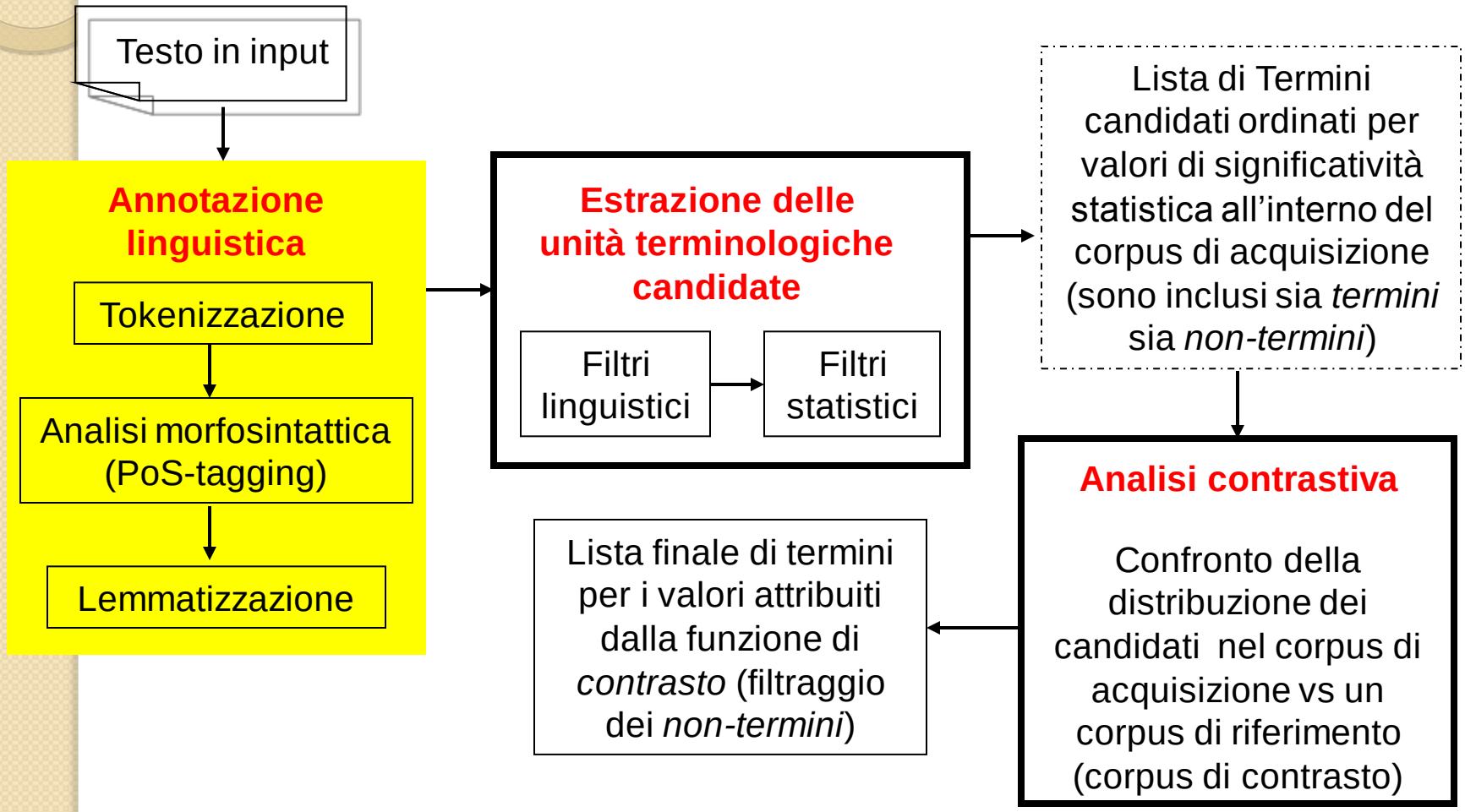
T2K: Terminology Extraction

- T2K usa un approccio multi-livello per l'estrazione dei termini



T2K: Terminology Extraction

- T2K usa un approccio multi-livello per l'estrazione dei termini



T2K: Terminology Extraction

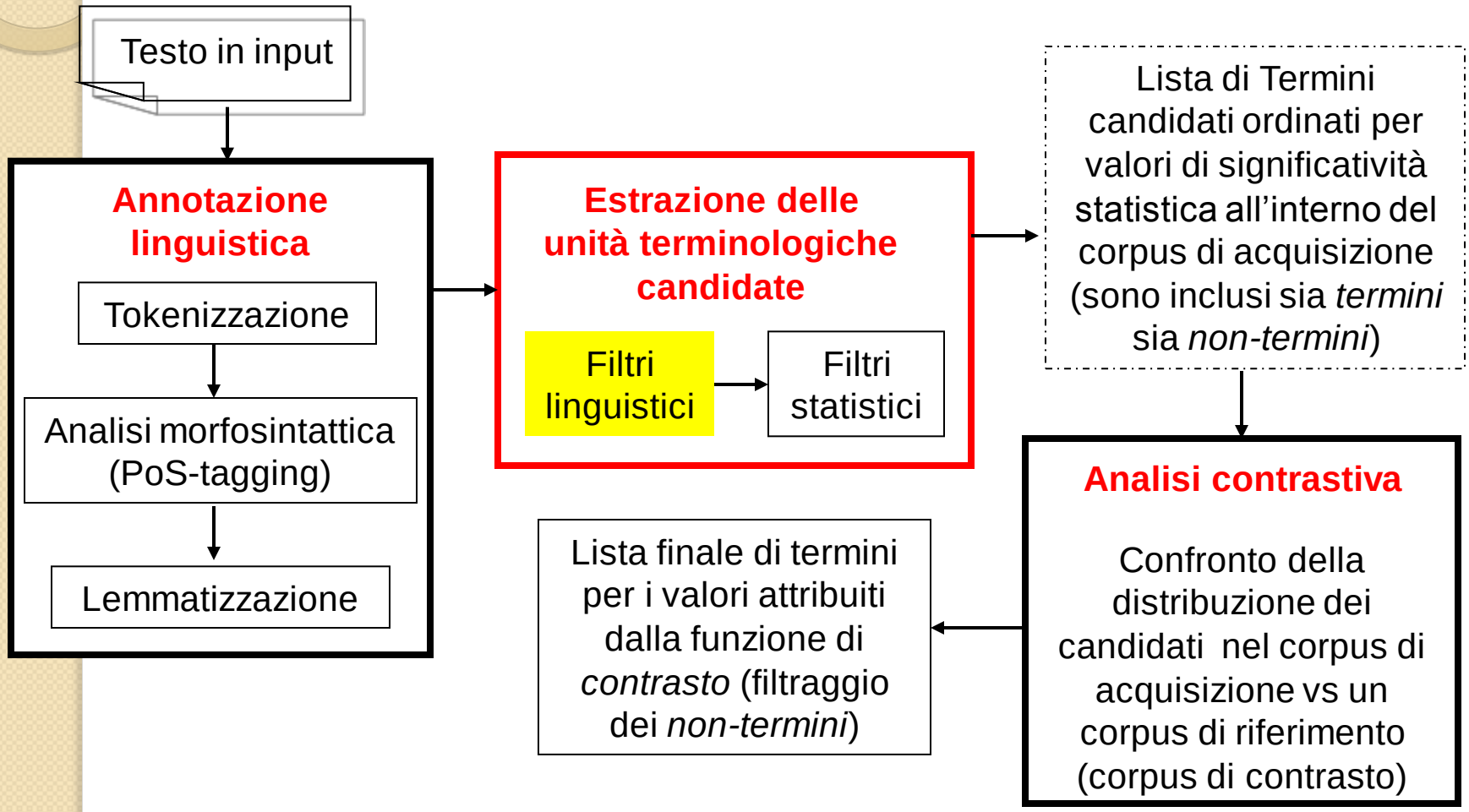
- Analisi linguistica fino al Part-Of-Speech tagging e Lemmatizzazione
  - E.g. Il piano nazionale di riduzione delle emissioni in nessun caso può esonerare un impianto dal rispetto della pertinente normativa comunitaria, compresa la direttiva 96/61/CE (*The national emission reduction plan may under no circumstances exempt a plant from the provisions laid down in relevant Community legislation, including inter alia Directive 96/61/EC*)

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
Il	il	R	RD	num=s|gen=m
piano	piano	S	S	num=s|gen=m
nazionale	nazionale	A	A	num=s|gen=n
di	di	E	E	_
riduzione	riduzione	S	S	num=s|gen=f
delle	di	E	EA	num=p|gen=f
emissioni	emissione	S	S	num=p|gen=f
in	in	E	E	_
nessun	nessun	D	DI	num=s|gen=m
caso	caso	S	S	num=s|gen=m
può	potere	V	VM	num=s|per=3|mod=i|ten=p
esonerare	esonerare	V	V	mod=f

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
un	un	R	RI	num=s|gen=m
impianto	impianto	S	S	num=s|gen=m
dal	da	E	EA	num=s|gen=m
rispetto	rispetto	S	S	num=s|gen=m
della	di	E	EA	num=s|gen=f
pertinente	pertinente	A	A	num=s|gen=n
normativa	normativa	S	S	num=s|gen=f
comunitaria	comunitario	A	A	num=s|gen=f
,	,	F	FF	_
compresa	comprendere	V	V	num=s|mod=p|gen=f
la	il	R	RD	num=s|gen=f
direttiva	direttiva	S	S	num=s|gen=f
96/61/CE.	96/61/CE.	S	SP	_

T2K: Terminology Extraction

- T2K usa un approccio multi-livello per l'estrazione dei termini



Estrazione di unità terminologiche candidate

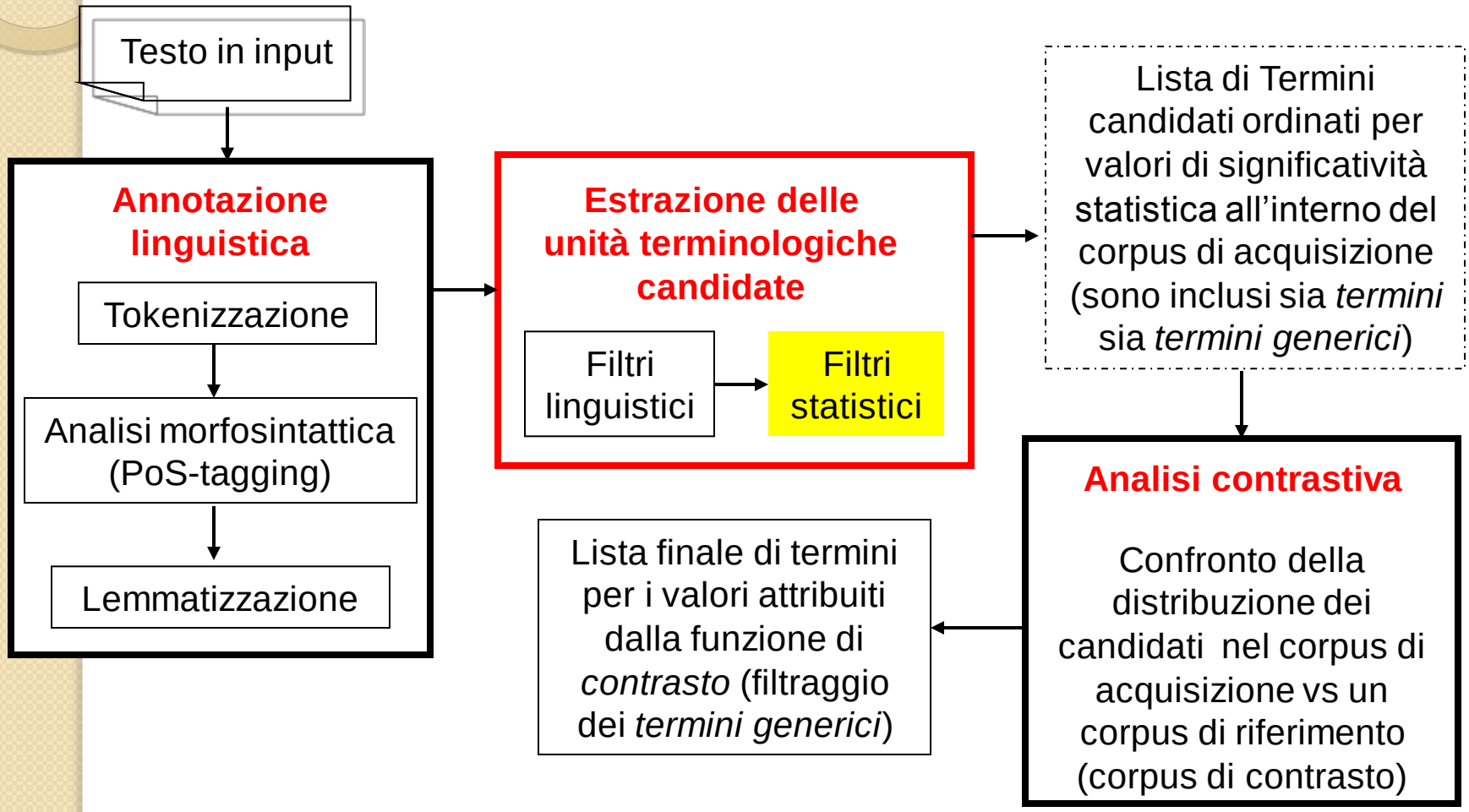
- Filtri linguistici
  - sostantivi (S), es. *impianto, direttiva*
  - sequenze di categorie morfosintattiche, quali
    - sostantivo+preposizione+sostantivo (S+E+S), es. *riduzione di emissione*
    - sostantivo+aggettivo (S+A), es. *piano nazionale, normativa comunitaria*

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
Il	il	R	RD	num=s|gen=m
piano	**piano**	**S**	**S**	**num=s|gen=m**
nazionale	**nazionale**	**A**	**A**	**num=s|gen=n**
di	di	E	E	_
riduzione	**riduzione**	**S**	**S**	**num=s|gen=f**
delle	**di**	**E**	**EA**	**num=p|gen=f**
emissioni	**emissione**	**S**	**S**	**num=p|gen=f**
in	in	E	E	_
nessun	nessun	D	DI	num=s|gen=m
caso	**caso**	**S**	**S**	**num=s|gen=m**
può	potere	V	VM	num=s|per=3|mod=i|ten=p
esonerare	esonerare	V	V	mod=f

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
un	un	R	RI	num=s|gen=m
impianto	**impianto**	**S**	**S**	**num=s|gen=m**
dal	**da**	**E**	**EA**	**num=s|gen=m**
rispetto	**rispetto**	**S**	**S**	**num=s|gen=m**
della	di	E	EA	num=s|gen=f
pertinente	pertinente	A	A	num=s|gen=n
normativa	**normativa**	**S**	**S**	**num=s|gen=f**
comunitaria	**comunitario**	**A**	**A**	**num=s|gen=f**
,	,	F	FF	_
compresa	comprendere	V	V	num=s|mod=p|gen=f
la	il	R	RD	num=s|gen=f
direttiva	**direttiva**	**S**	**S**	**num=s|gen=f**
96/61/CE.	96/61/CE.	S	SP	_

T2K: Terminology Extraction

- T2K usa un approccio multi-livello per l'estrazione dei termini



Estrazione di unità terminologiche candidate

- Filtri statistici
  - C-NC Value (Frantzi & Ananiadou 1999) per determinare la probabilità di un'unità poliremática di essere un termine
    - vengono eliminati *non-termini*, es: ***impianto dal rispetto***

filtri statistici (ranking)

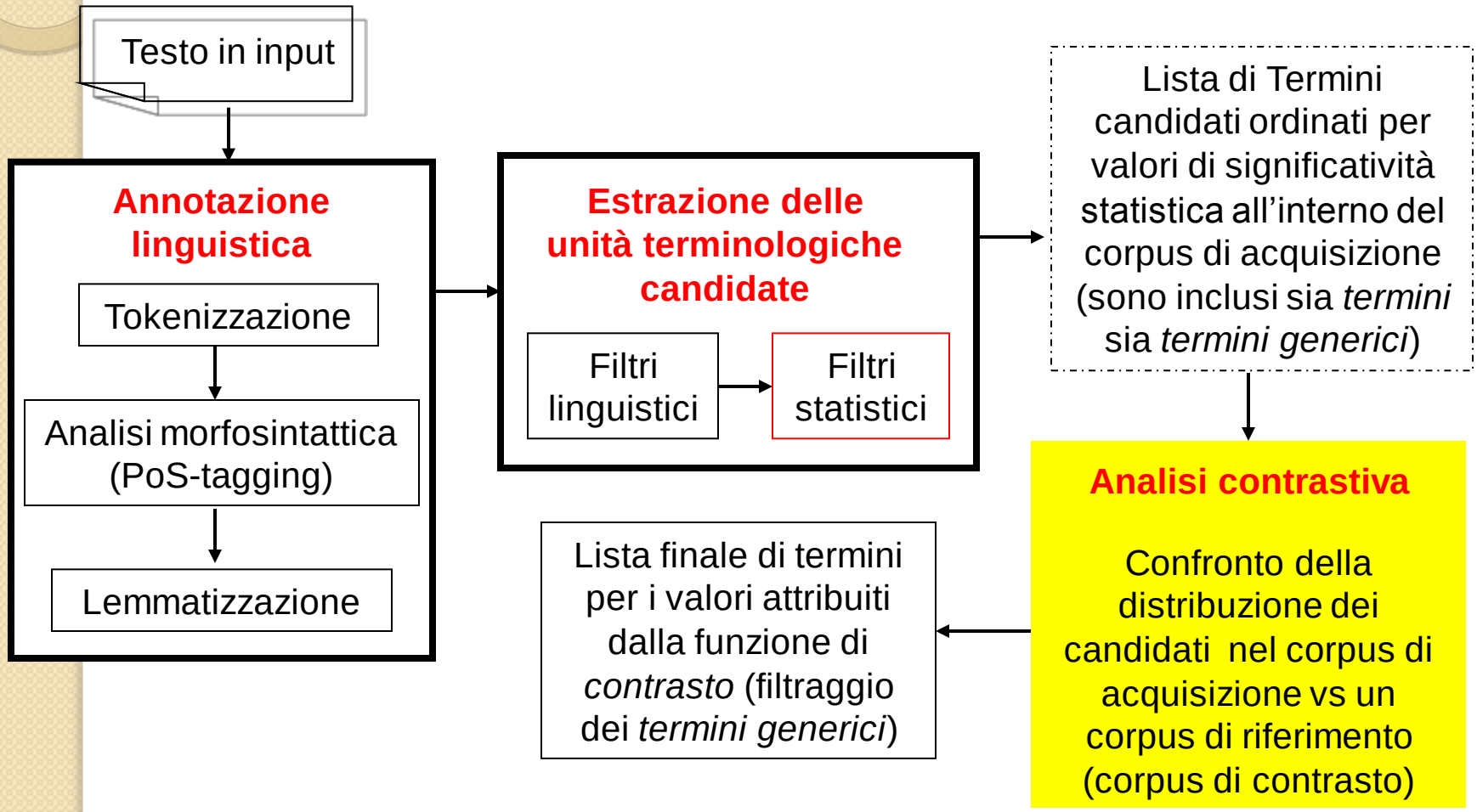
autorità competente	236.120380272
piano nazionale	113.117778156
riduzione delle emissioni	108.219717591
direttiva	105.211324357
valore limite di emissione	103.436822534
destinatario della decisione	87.2457638653
limite di emissione	86.9062873351
sostanza pericolosa	84.8930693328
caso	*37.5790064648*
anno precedente	23.934467506
danno ambientale	37.4660023032

Risultati dei filtri statistici:

Termini generici, **termini del dominio legale**, **termini specifici del dominio regolato** **mischiati insieme**

T2K: Terminology Extraction

- T2K usa un approccio multi-livello per l'estrazione dei termini



T2K: applicazione del contrasto

Funzione di Contrasto

filtri statistici (ranking)

autorità competente	236.120380272
piano nazionale	113.117778156
riduzione delle emissioni	108.219717591
direttiva	105.211324357
valore limite di emissione	103.436822534
destinatario della decisione	87.2457638653
limite di emissione	86.9062873351
sostanza pericolosa	84.8930693328
caso	*37.5790064648*
anno precedente	23.934467506
danno ambientale	37.4660023032

Contrasta la lista di termini con i termini estratti da un altro dominio (in questo caso un dominio diverso di direttive Europee)

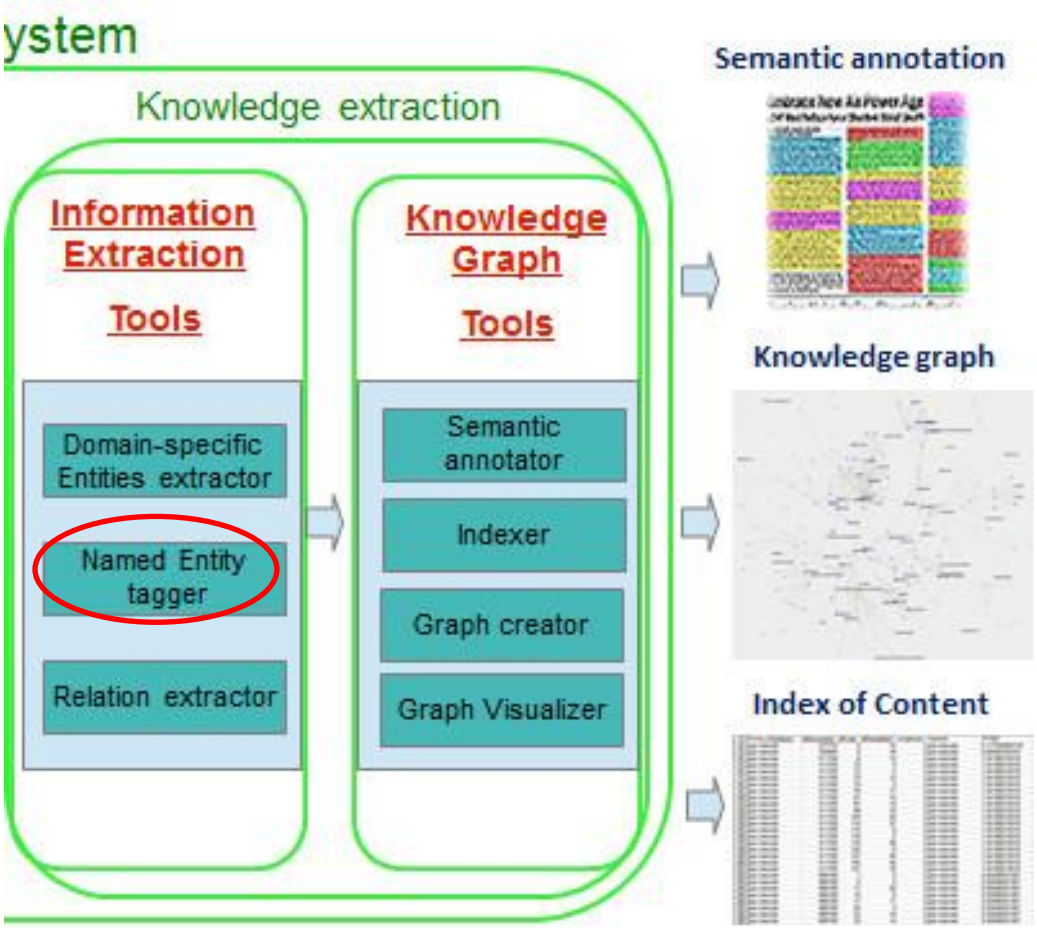
Risultato della funzione di Contrasto:

I **termini del dominio giuridico** e quelli ***generici (o vuoti)*** vengono separati dai **termini dello specifico dominio** trattato (ambientale)

Lista finale

riduzione delle emissioni	100
valore limite di emissione	98.5
limite di emissione	98.3
sostanza pericolosa	84.89
danno ambientale	84
.....	
autorità competente	30
piano nazionale	29
Direttiva	28.45
destinatario della decisione	28.3
caso	*27*
anno precedente	25

T2K: Estrazione e classificazione di entità nominate (NERC)



Estrazione e classificazione di entità nominate (NERC)

T2K utilizza algoritmi di apprendimento supervisionato per la risoluzione di questo compito. NERC come compito di classificazione : *assegnare* (o meglio classificare) *ogni token all'interno della frase a una delle possibili classi di output* (es 5 classi per l'italiano):

PERSONA, LUOGO, ORGANIZZAZIONE, LUOGO GEOPOLITICO, *ALTRO*

Giacomo	PERSONA
Leopardi	PERSONA
scrisse	*ALTRO*
L'	*ALTRO*
Infinito	*ALTRO*
a	*ALTRO*
Recanati	LUOGO GEOPOLITICO
.	ALTRO

Estrazione e classificazione di entità nominate (NERC)

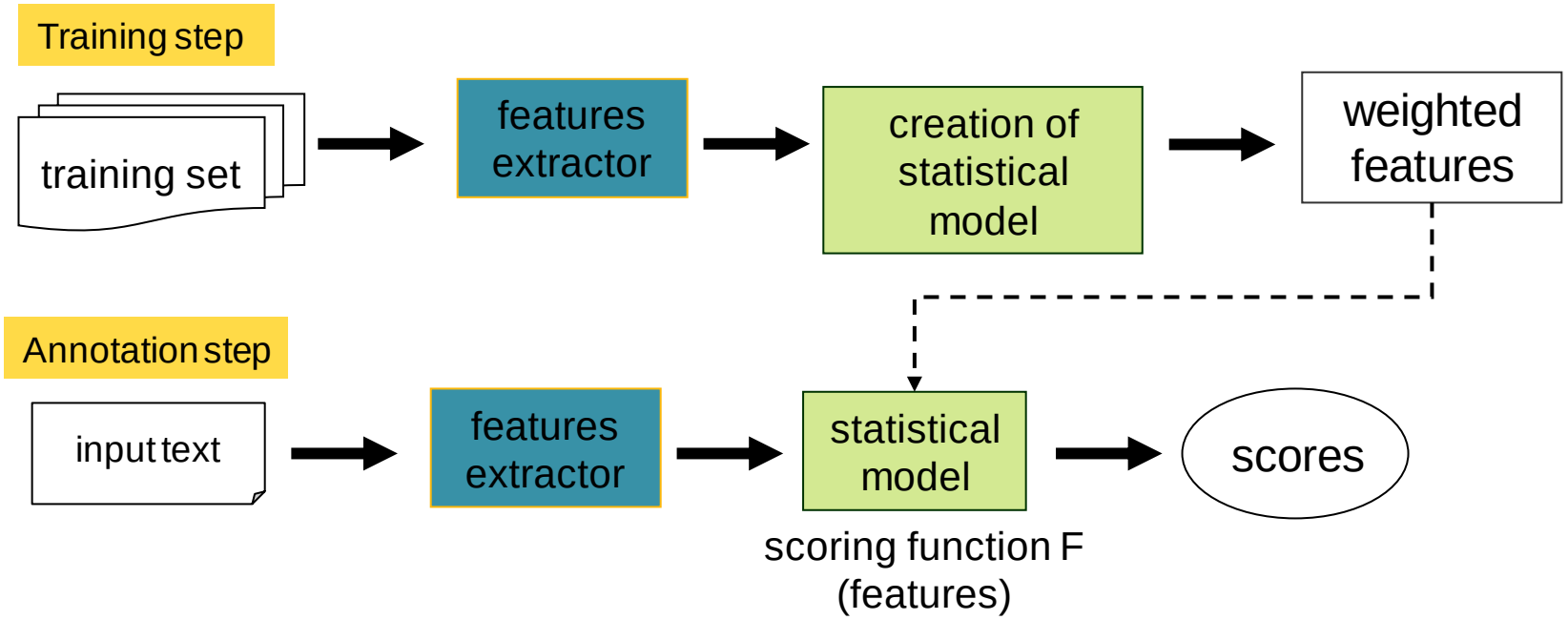
Alcune parole identificano insieme una unica entità nominata, per definire questa proprietà T2K utilizza il formato standard: BIO-format *(Begin, Inside, Otside).*

Il numero di classi aumenta: **B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, B-GPE, I-GPE, *O***

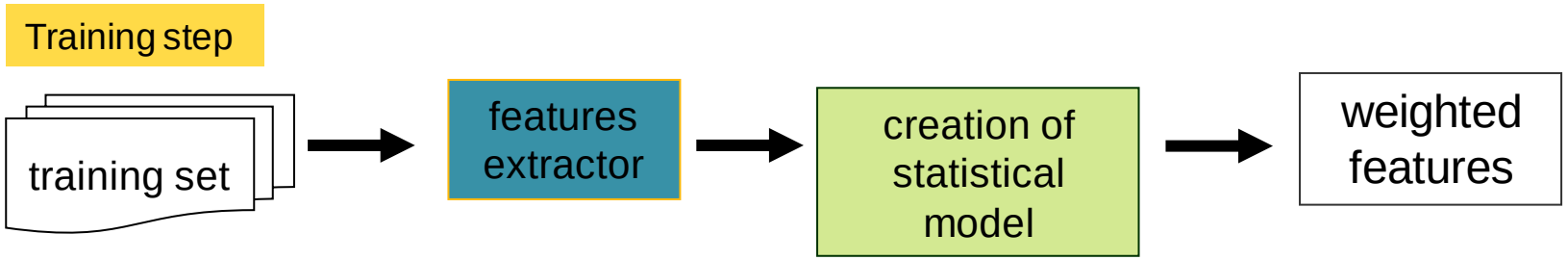
Giacomo	*B*-PER
Leopardi	*I*-PER
scrisse	*O*
L'	*O*
Infinito	*O*
a	*O*
Recanati	*B-GPE*
.	O

Algoritmo di Apprendimento Supervisionato

Il funzionamento di un algoritmo basato sull'apprendimento supervisionato può essere diviso in due fasi:



Algoritmo di Apprendimento Supervisionato: addestramento



training set: corpus di esempi annotati: coppie *(input, output)*

feature: caratteristiche estratte dall'input (training set)

Modello statistico: insieme di coppie (feature, *peso*), dove il *peso* è stato calcolato dall'algoritmo di apprendimento

Algoritmo di Apprendimento Supervisionato: estrazione features

Esempio di frase annotata con le entità nominate (frase presente nel training set):

(*La* , *O*)

(*Roma* , *B-ORGANIZZAZIONE*)

(*ha* , *O*)

(*vinto* , *O*)

(*la* , *O*)

(*partita* , *O*)

(*a* , *O*)

(*Milano* , *B-GPE*)

Feature attive per la classe B-ORGANIZZAZIONE

Features Locali:
Suffisso=**Ro**
Prefisso=**ma**
Forma=**Roma**
Lunghezza=**4**

Algoritmo di Apprendimento Supervisionato: estrazione features

Esempio di frase annotata con le entità nominate (frase presente nel training set):

(*La* , *O*)

(*Roma* , *B-ORGANIZZAZIONE*)

(*ha* , *O*)

(*vinto* , *O*)

(*la* , *O*)

(*partita* , *O*)

(*a* , *O*)

(*Milano* , *B-LUOGO*)

Feature attive per la classe B-ORGANIZZAZIONE

Features Contestuali:
Token(-1)=**La**
Token(+1)=**ha**
Token(+2)=**vinto**
…

Algoritmo di Apprendimento Supervisionato: estrazione features

Esempio di frase annotata con le entità nominate (frase presente nel training set):

(*La* , *O*)

(*Roma* , *B-ORGANIZZAZIONE*)

(*ha* , *O*)

(*vinto* , *O*)

(*la* , *O*)

(*partita* , *O*)

(*a* , *O*)

(*Milano* , *B-LUOGO*)

Feature attive per la classe B-ORGANIZZAZIONE

Features Globali:
Tipo_Del_documento=Sportivo

Algoritmo di Apprendimento Supervisionato: estrazione features

Il processo di estrazione delle feature restituisce per ogni coppia *(input, output)* la lista delle feature attive in quel contesto per la classe *output*. *Nell'esempio precedente:*

Features Attive per B-ORG:
***Suffisso*=Ro**
***Prefisso*=ma**
***Forma*=Roma**
***Lunghezza*=4**
Token(-1)=**La**
Token(+1)=**ha**
Token(+2)=**vinto**
Tipo_documento=Sport

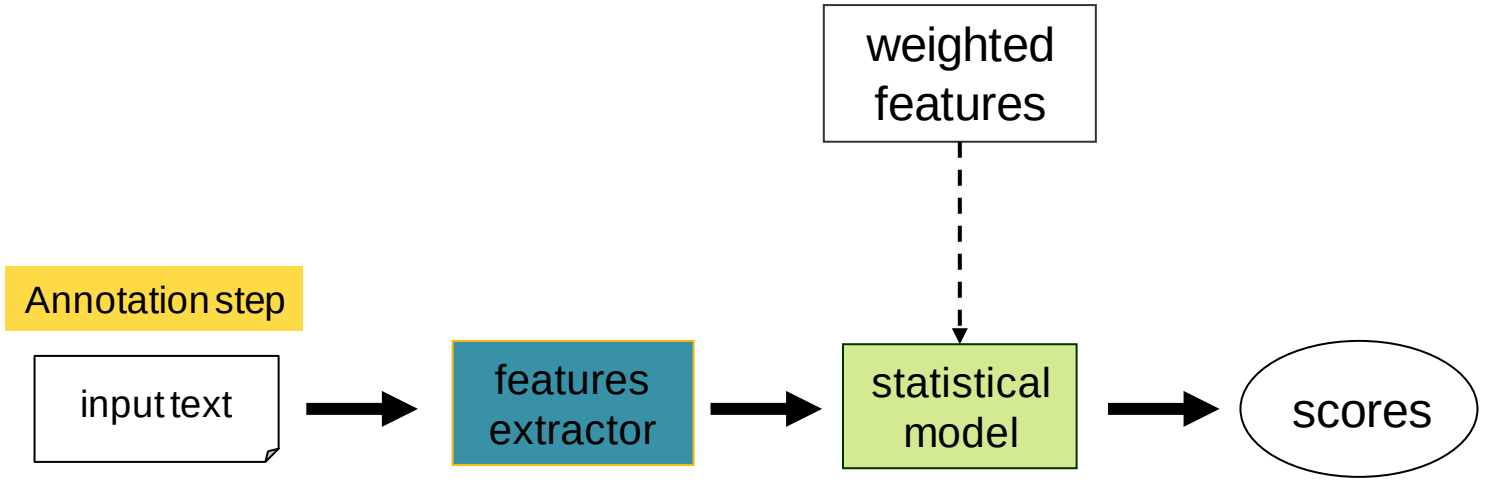
Queste feature si aggiungono a quelle già estratte per la stessa categoria (B-ORG) in altri eventi annotati all'interno del corpus.

Algoritmo di Apprendimento Supervisionato

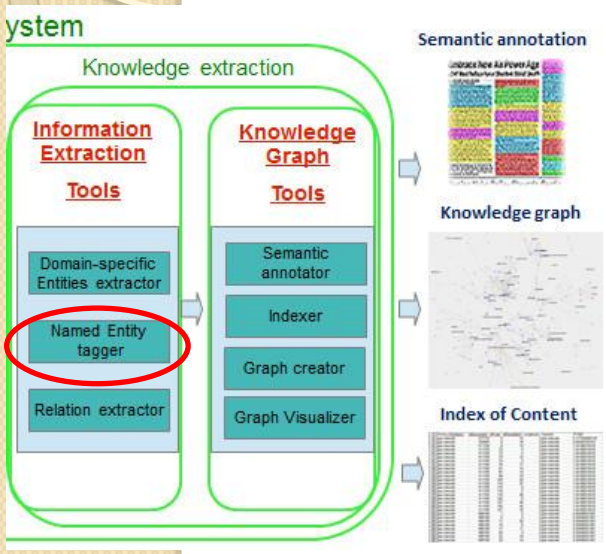
Alla fine del processo di estrazione su tutto il corpus, le feature vengono *pesate* dall'algoritmo di apprendimento automatico.

I *pesi* indicano la "forza" della feature nell'indicare una certa *classe* come possibile *output e* possono essere visti come i parametri della funzione obiettivo e come il *modello della lingua* che il sistema di addestramento crea nella fase di apprendimento.

Algoritmo di Apprendimento Supervisionato: analisi



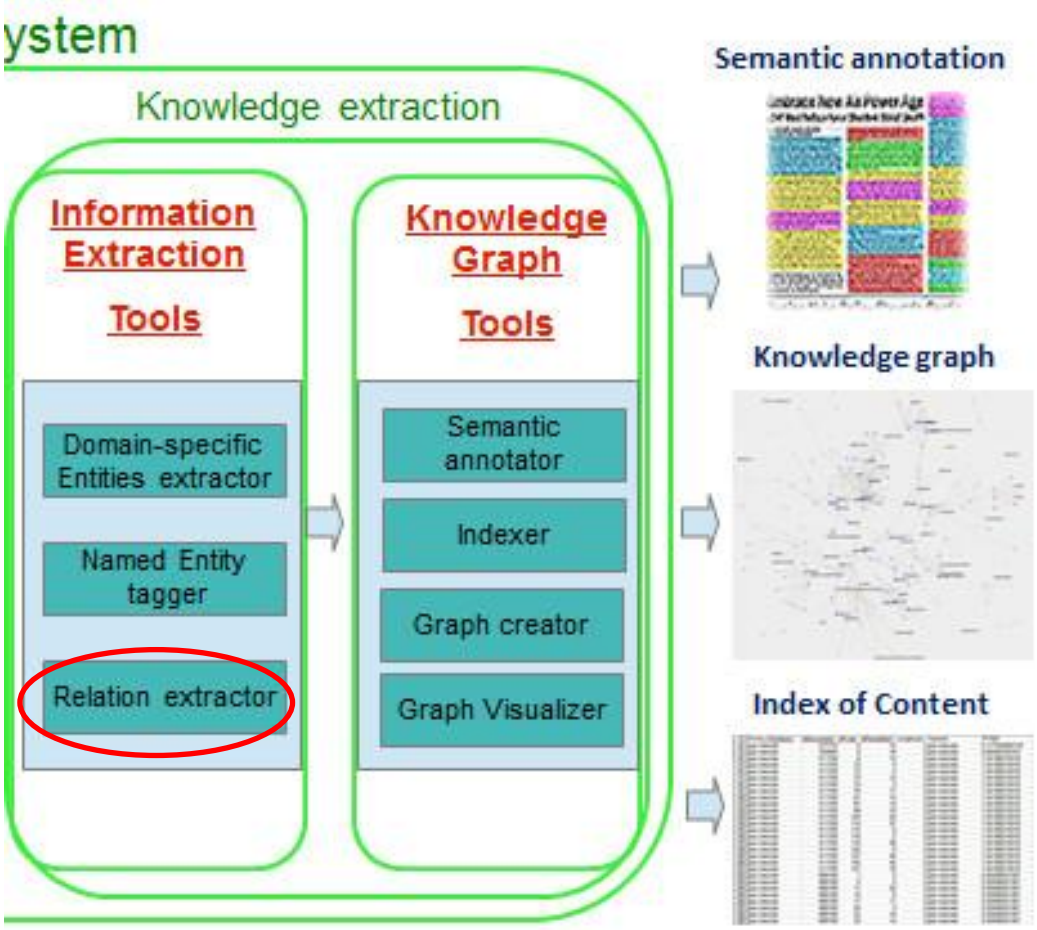
T2K: NERC



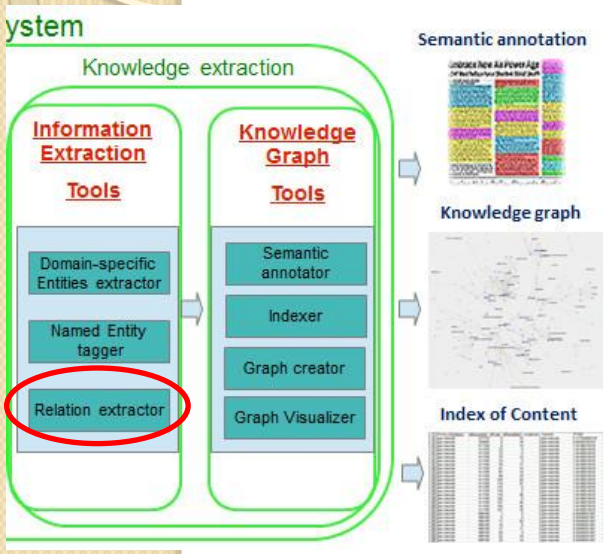
Input corpus: collezione di Direttive Europee in materia ambientale

Entity	Class	Frequency	Frequency (%)	Class Frequency (%)
Commission	Organization	980.0	5.55	7.35
EC	Organization	362.0	2.05	2.72
EEC	Organization	341.0	1.93	2.56
European Union	Organization	311.0	1.76	2.33
European Parliament	Organization	184.0	1.04	1.38
EU	Organization	142.0	0.80	1.07
Brussels	Location	118.0	0.67	3.61
European Community	Organization	110.0	0.62	0.83
Council of Europe	Organization	77.0	0.44	0.58
Europe	Location	72.0	0.41	2.20
Rome	Location	66.0	0.37	2.02
Consultative Committee	Organization	65.0	0.37	0.49
Council	Organization	62.0	0.35	0.47
Euratom	Organization	62.0	0.35	0.47
Schengen	Location	59.0	0.33	1.80
Romania	Location	58.0	0.33	1.77
Luxembourg	Location	50.0	0.28	1.53
Association Council	Organization	45.0	0.25	0.34
Management Board	Organization	45.0	0.25	0.34
Bulgaria	Location	45.0	0.25	1.38
Schengen Information System	Organization	43.0	0.24	0.32
Stabilisation and Association Council	Organization	42.0	0.24	0.32
EUROPEAN UNION	Organization	40.0	0.23	0.30
European Council	Organization	37.0	0.21	0.28
Ireland	Location	34.0	0.19	1.04
France	Location	33.0	0.19	1.01
Northern Ireland	Location	30.0	0.17	0.92
Republic of Cyprus	Location	30.0	0.17	0.92

T2K: Estrazione di relazioni



T2K: Estrazione di relazioni

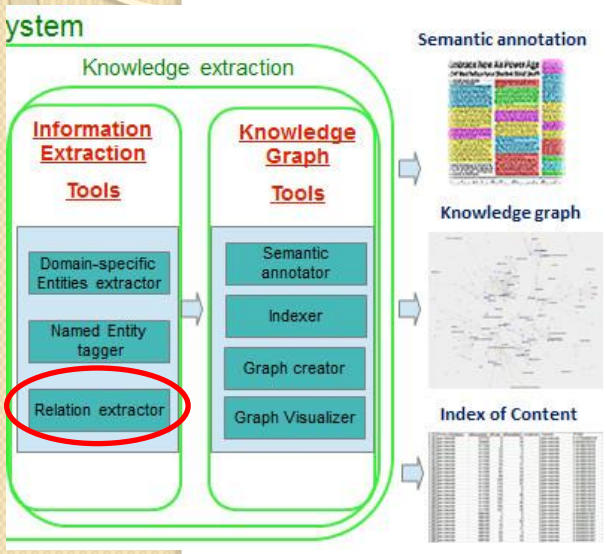


E.g.: termini in relazione con ***imaging cerebrale*** in testi di giurisprudenza penale

imaging cerebrale (*brain imaging*)	
genetica molecolare (*molecular genetics*)	quadro clinico (*medical case*)
difesa (*defense*)	comportamenti illeciti (*illegal behaviours*)
valutazione (*evaluation*)	nesso causale (*causal relationship*)
colloqui clinici (*clinical interviews*)	apporto tecnico (*technical contribution*)
emergenze psichiatriche (*psychiatric emergencies*)	sfera psichica (*psychic sphere*)
accertamenti psichiatrici (*psychiatric inspections*)	imputata (*defendant*)

Input corpus:
collezione di sentenze penali italiane nelle quali si fa uso della prova neuroscentifica

T2K: Estrazione di relazioni



E.g.: termini in relazione con ***imaging cerebrale*** in testi di giurisprudenza

imaging cerebrale (*brain imaging*)	
genetica molecolare (*molecular genetics*)	quadro clinico (*medical case*)
difesa (*defense*)	comportamenti illeciti (*illegal behaviours*)
valutazione (*evaluation*)	nesso causale (*causal relationship*)
colloqui clinici (*clinical interviews*)	apporto tecnico (*technical contribution*)
emergenze psichiatriche (*psychiatric emergencies*)	sfera psichica (*psychic sphere*)
accertamenti psichiatrici (*psychiatric inspections*)	imputata (*defendant*)

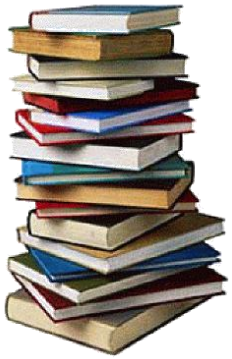
e.g. contesto

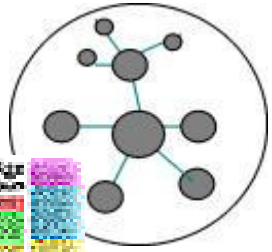
Input corpus: collezione di sentenze penali italiane nelle quali si fa uso della prova neuroscentifica

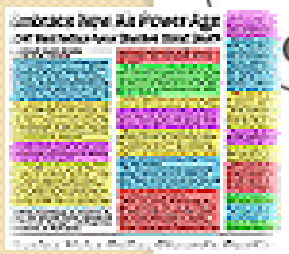
Sia le emergenze psichiatriche , completate dalle risultanze dell' imaging cerebrale e di genetica molecolare , che quelle processuali consentono di rilevare gravi segni di disfunzionalità psichica , eterogenei ma convergenti nell' indicare un nesso causale tra i disturbi dell' imputata ed i suoi comportamenti illeciti

Dal testo alla conoscenza: l'approccio generale

Processo incrementale di annotazione-acquisizione-annotazione:
knowledge acquired from linguistically-annotated texts is projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted







Textual content
(implicit knowledge)

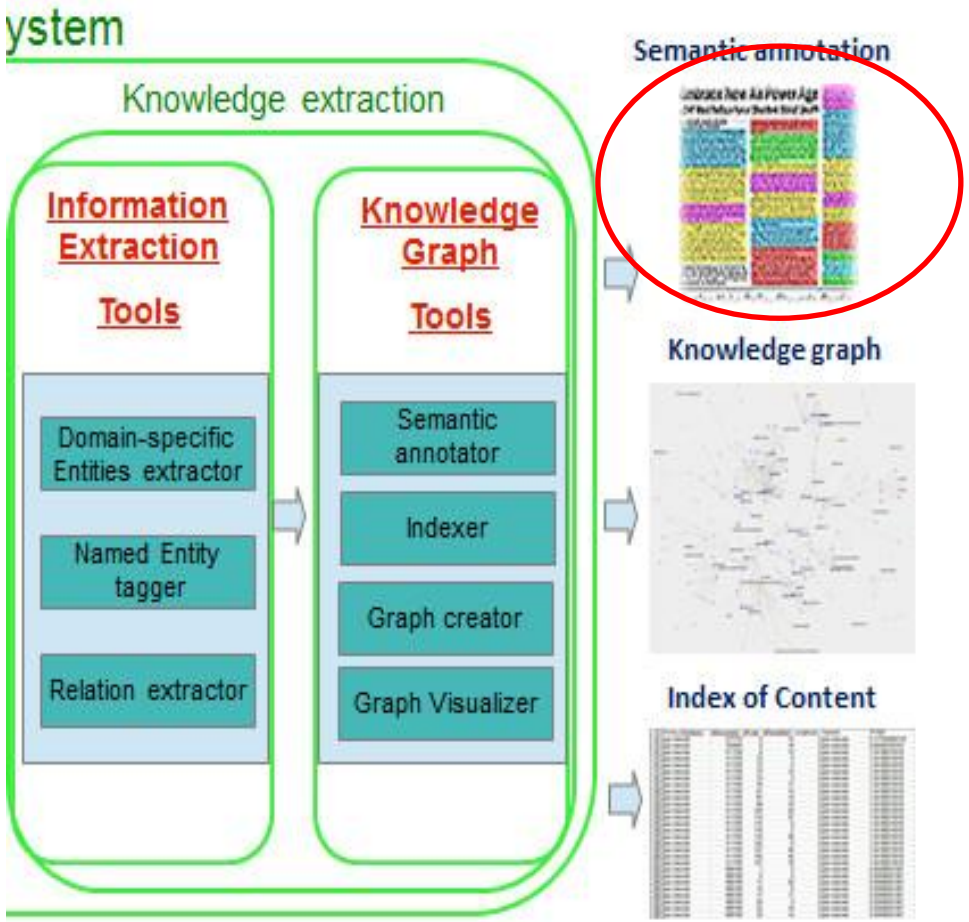
Dynamic content structuring

Linguistic annotation

Knowledge extraction

Structured knowledge
(explicit knowledge)

T2K: Annotazione Semantica



T2K: Annotazione Semantica

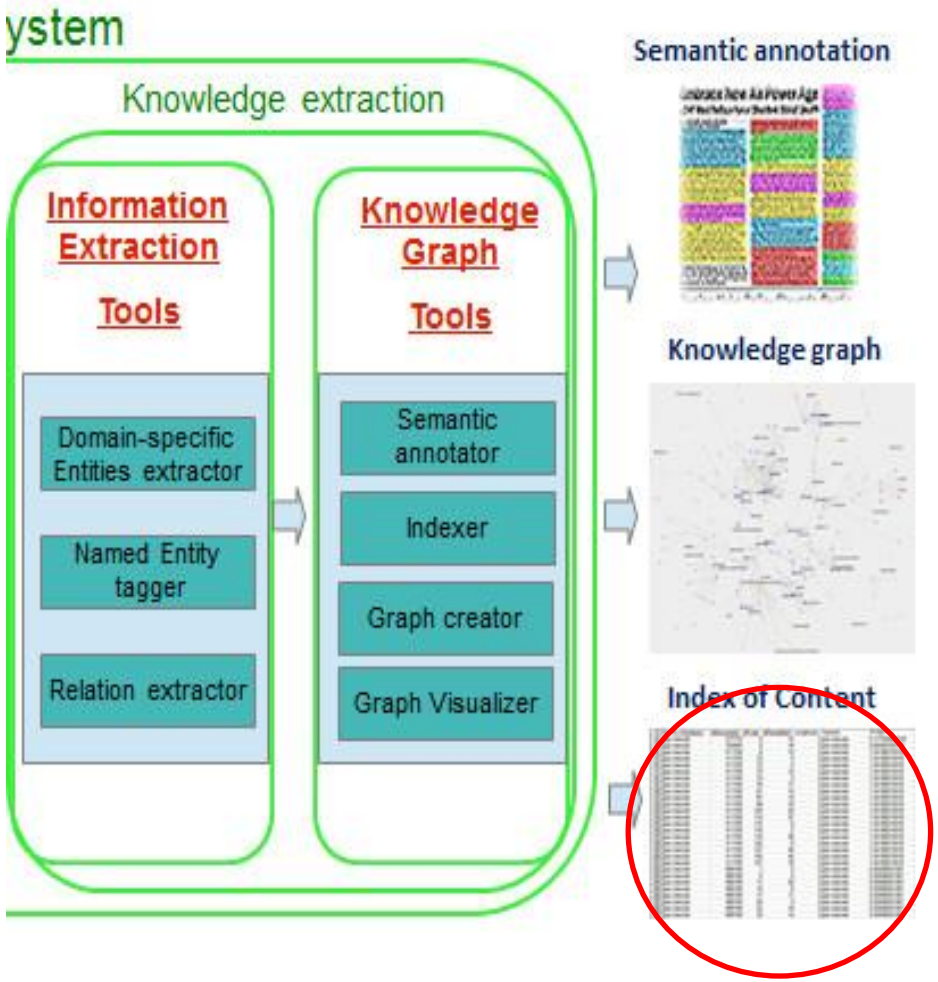
La conoscenza estratta (es. termini, entità nominate) viene riproiettata sul corpus

e.g.

Input corpus: collezione di sentenze in materia di *responsabilità dello stato*

La sentenza ritiene azionato, pur in assenza di espressa qualificazione in tal senso nell'atto introduttivo del giudizio, il **diritto al risarcimento del danno**, ex art. 2043 c.c., per **violazione dell'obbligo** dello Stato di dare attuazione alle **direttive comunitarie** che imponevano di remunerare adeguatamente il medico per la frequenza di un **corso di specializzazione**; considera comprovato, in assenza di contestazioni specifiche, che il C. avesse superato il corso di formazione quadriennale, come da attestazione del 5.11.1992, con frequenza a tempo pieno e senza svolgimento di attività libero-professionale; dichiara inammissibile l'eccezione di **prescrizione quinquennale** sollevata dall'amministrazione ed accolta dal primo giudice, sul rilievo che era stata formulata, senza le necessarie allegazioni in fatto e diritto, con riferimento all'art. 2948 c.c., n. 4, in termini, quindi, non pertinenti al rapporto giuridico dedotto in giudizio, atteso che non si trattava di rapporto di impiego pubblico (prospettazione su cui si fondava il **difetto di giurisdizione** ordinaria, eccepito dall'amministrazione in primo grado) e di **responsabilità contrattuale**; liquida il risarcimento nell'importo di L. 13.000.000 annue (Euro 6.713,93) secondo il parametro fornito dalla L. n. 370 del 1999, art. 1, comma 1 (**borsa di studio** annuale per i medici ammessi presso le università alle **scuole di specializzazione** in medicina dall'anno accademico 1983-1984 all'anno accademico 1990-1991, in attuazione di giudicati amministrativi), con l'aggiunta della rivalutazione monetaria e degli interessi legali dalla maturazione del credito, fissata alla data del 5 novembre 1992.

T2K: Indicizzazione



T2K: Indicizzazione

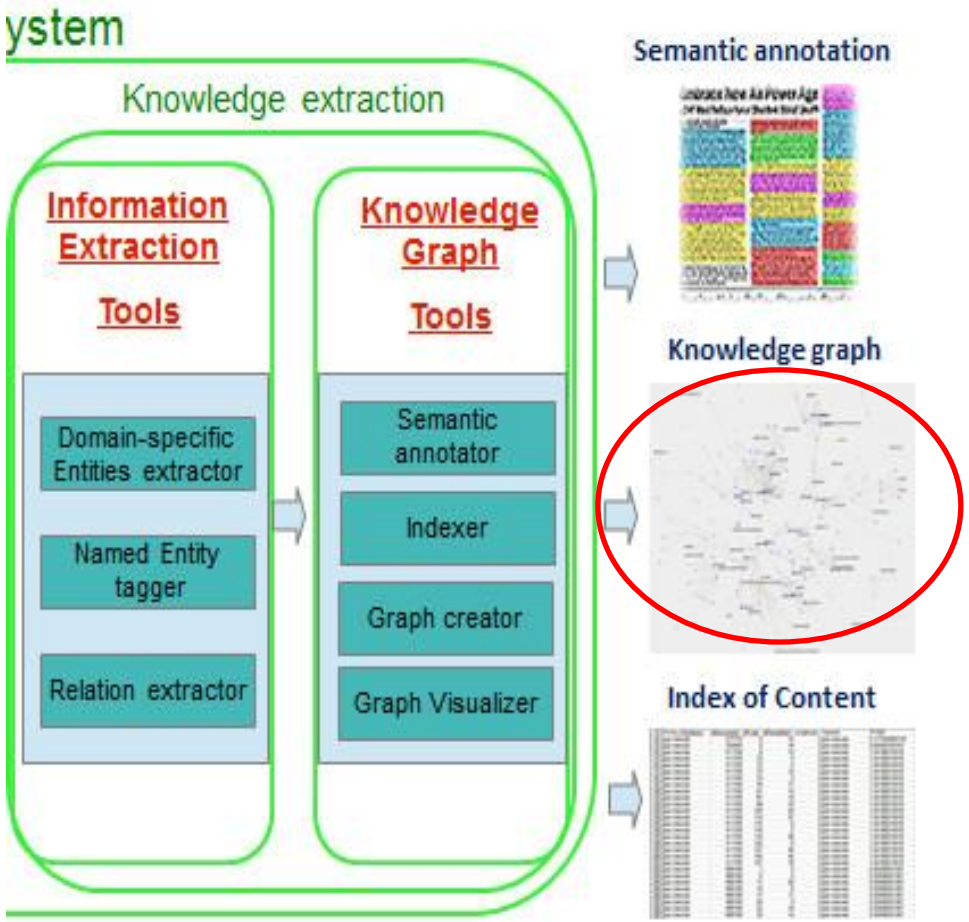
Input corpus:
orpus di relazioni sulle eco-mafie della Direzione Nazionale Antimafia

La conoscenza estratta (es. termini, entità nominate) viene utilizzata per indicizzare i documenti del corpus

es.

	Forma_Prototipica	idDocumento	idFrase	idParolaStart	Lunghezza	Variante	TF*IDF
2	area lombarda	2008/brescia.txt	12	75	2	area lombarda	0.012627727456
3	area lombarda	2008/brescia.txt	119	10	2	area lombarda	0.012627727456
4	immigrazione clandestina	2008/trieste.txt	5	41	2	immigrazione clandestina	0.0250679566129
5	immigrazione clandestina	2008/brescia.txt	101	47	2	immigrazione clandestina	0.012966184455
6	immigrazione clandestina	2008/brescia.txt	117	38	2	immigrazione clandestina	0.012966184455
7	immigrazione clandestina	2008/brescia.txt	118	49	2	immigrazione clandestina	0.012966184455
8	autorità giudiziarie	2008/trieste.txt	19	22	2	autorità giudiziarie	0.0135155036036
9	autorità giudiziarie	2008/trento.txt	8	58	2	autorità giudiziarie	0.0147441857494
10	autorità giudiziarie	2008/milano.txt	18	59	2	autorità giudiziarie	0.00730567762357
11	autorità giudiziarie	2008/milano.txt	24	78	2	autorità giudiziaria	0.00730567762357
12	autorità giudiziarie	2008/milano.txt	79	19	2	autorità giudiziaria	0.00730567762357
13	autorità giudiziarie	2008/genova.txt	31	48	2	autorità giudiziarie	0.0065397598082
14	reato associativo	2008/brescia.txt	45	39	2	reato associativo	0.0189415911839
15	reato associativo	2008/brescia.txt	59	7	2	reato associativo	0.0189415911839
16	reato associativo	2008/brescia.txt	101	39	2	reati associativi	0.0189415911839
17	traffici di droga	2008/trieste.txt	15	34	3	traffici di droga	0.0135155036036
18	traffici di droga	2008/torino.txt	23	10	3	traffico di droga	0.0159005924748
19	traffici di droga	2008/milano.txt	55	17	3	traffico di droga	0.00487045174905
20	traffici di droga	2008/milano.txt	59	42	3	traffici di droga	0.00487045174905
21	traffici di droga	2008/brescia.txt	41	14	3	traffici di droga	0.002330259242

T2K: Organizzazione della conoscenza



T2K: Organizzazione della conoscenza

Input corpus:
libri di storia dell'arte

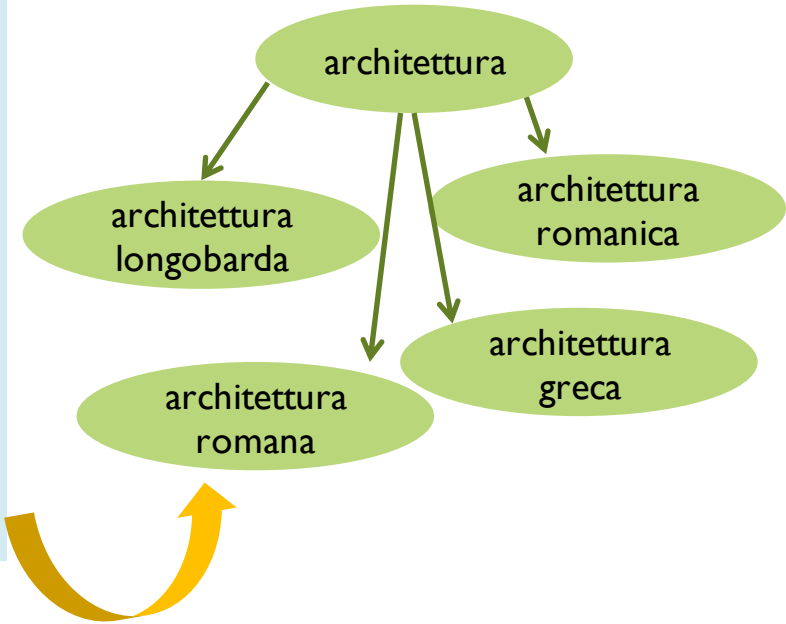
Persone
Giotto
Cimabue
Giovanni Pisano
Simone Martini
Arnolfo di Cambio
Ambrogio Lorenzetti
Cennino Cennini
Dante
Duccio di Buoninsegna
…

Luoghi
Roma
Italia
Assisi
Siena
Firenze
Pisa
Padova
Italia settentrionale
Saint-Denis
…

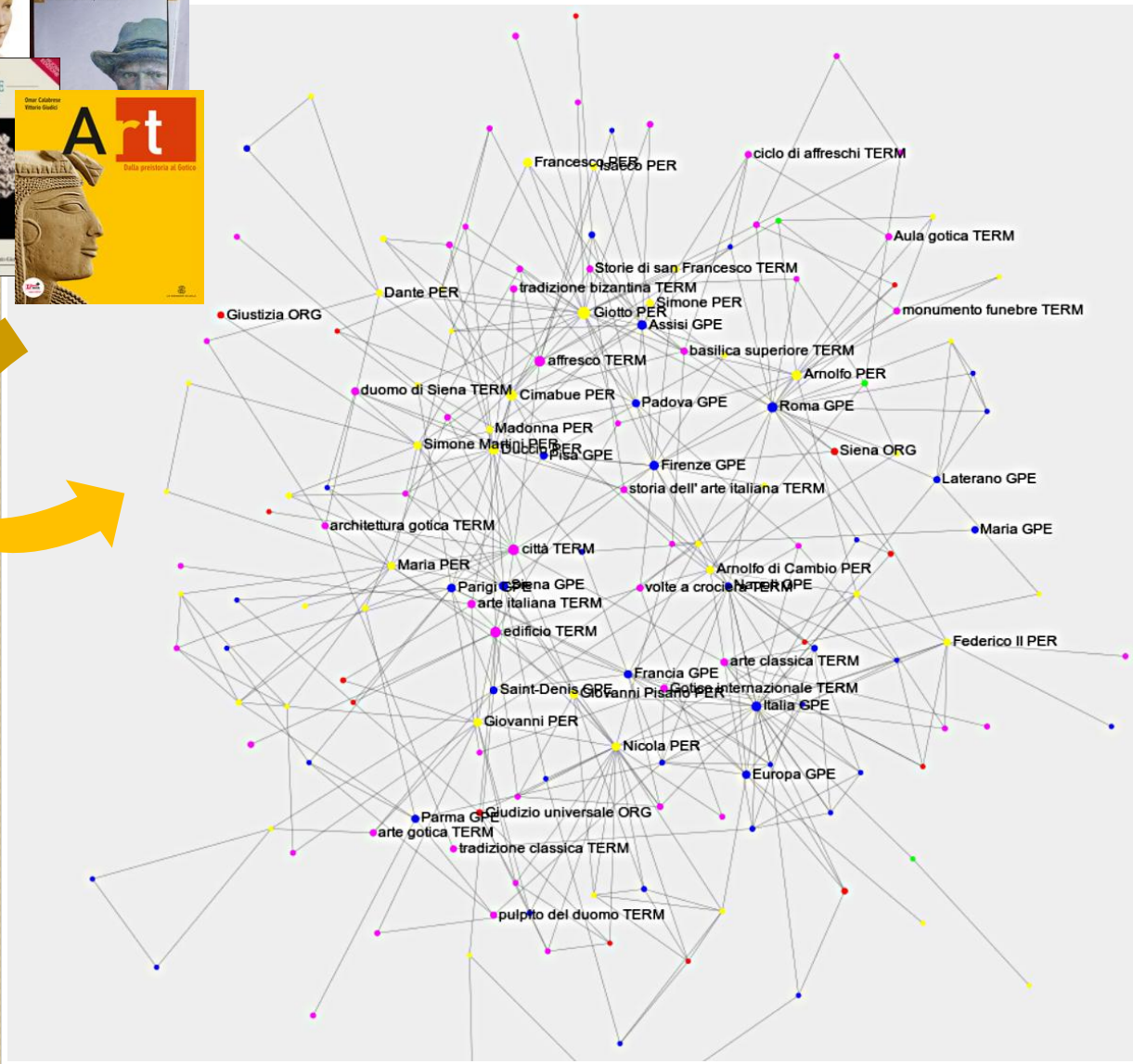
Organizzazioni
Sacro Romano Impero
Metropolitan Museum
Musée de Cluny
Collezione Salini
Museo Provinciale
…

Termini rilevanti
edificio
affresco
città
duomo di Siena
arte italiana
colonne
Giudizio Universale
storie di San Francesco
arte classica
architettura
gotico internazionale
ciclo di affreschi
pulpito del duomo
volte a crociera
tradizione bizantina
basilica superiore
…

Organizzazione tassonomica



T2K: Organizzazione della conoscenza



Display Options

Entity Minimum Frequency:

2

Relation Minimum Value:

1

Entity selection

Distance Metrics:

Log-likelihood

Submit

Search on the graph:

Type something

Many labels - Few labels:

Dante PER

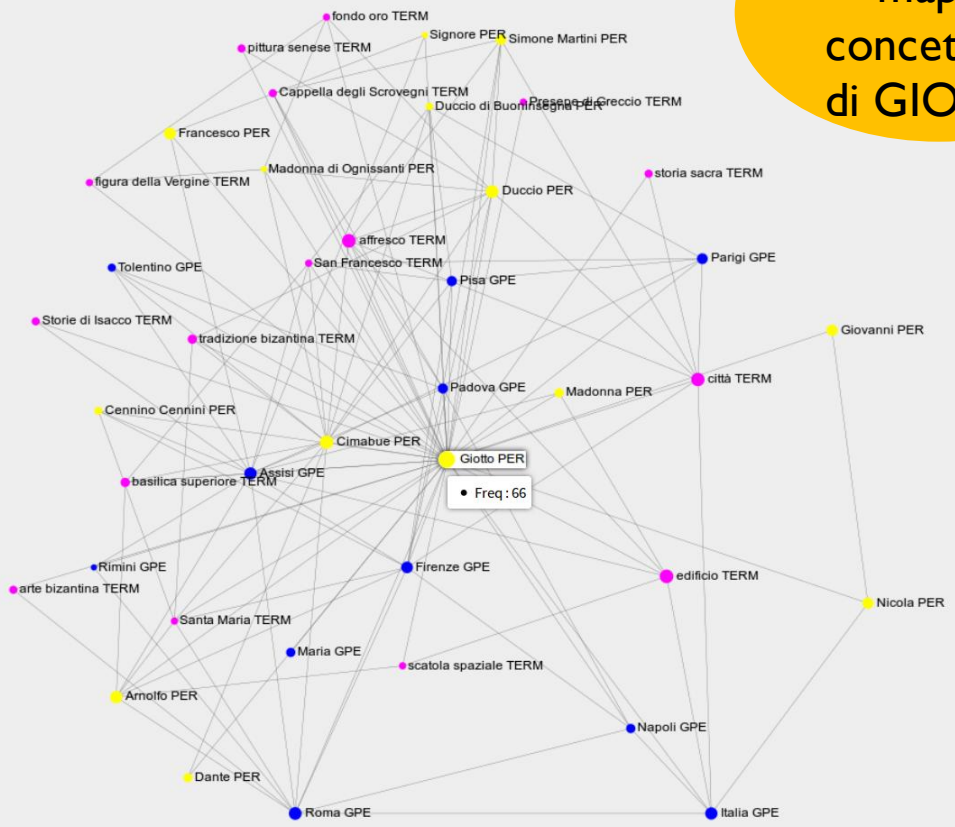
Node	Weight
Federico PER	7.24664172627
pittura di Giotto TERM	7.24664172627
Capua GPE	6.57417253588
architettura gotica TERM	6.0759627691
Giotto PER	5.03333778935
Cimabue PER	2.75404614865

Showing 1 to 6 of 6 entries

Previous Next

T2K: accesso al contenuto

Verso la mappa concettuale di GIOTTO



Display Options

Entity Minimum Frequency: 2

Relation Minimum Value: 1

Entity selection

Distance Metrics: Frequency

Submit

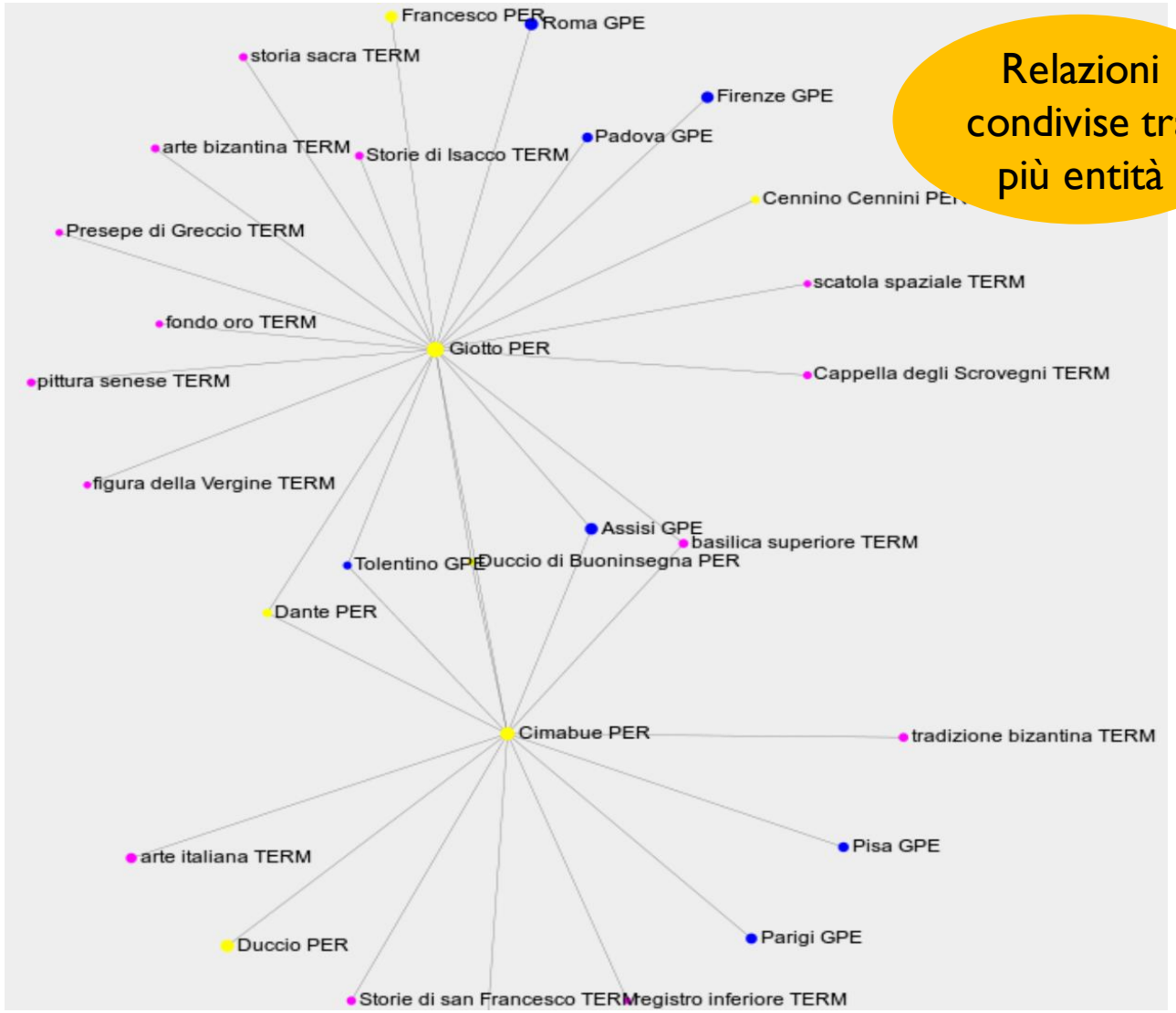
Search on the graph: Type something

Many labels - Few labels:

Giotto PER

Node	Weight
Cimabue PER	7.0
Assisi GPE	5.0
Firenze GPE	5.0
Padova GPE	4.0
Cennino Cennini PER	3.0
Roma GPE	3.0
Cappella degli Scrovegni TERM	3.0
San Francesco TERM	2.0
Duccio PER	2.0
Francesco PER	2.0
edificio TERM	2.0

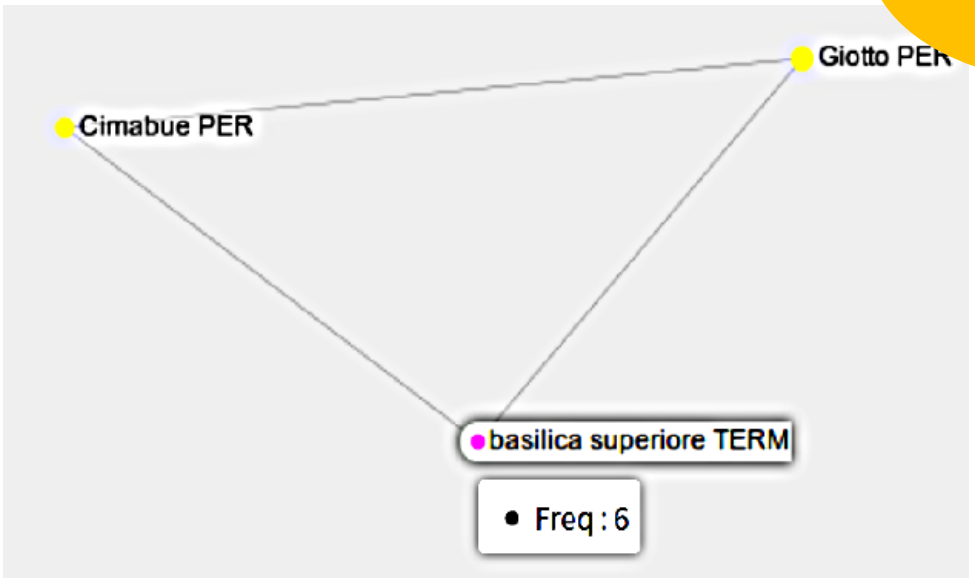
T2K: accesso al contenuto



Relazioni condivise tra più entità

T2K: accesso al contenuto

Relazioni condivise tra due entità



Con queste premesse fondamentali, Cimabue affronta negli anni Ottanta la decorazione del capocroce della basilica superiore di San Francesco ad Assisi, dove lavorerà di lì a poco il giovane Giotto, impegnato nell'esecuzione delle storie del santo.

Text-to-Knowledge (T2K) a lavoro

- T2K è utilizzato all'interno di numerosi progetti:
  - *Legal Text Mining: building semantic networks to support advanced queries in legal textual corpora* (JURNET)
  - *iSLe – intelligent Semantic Liquid eBook*
  - *INMOTO: INformation and MObility for Tourism*
  - Analisi di documentazione tecnica, come brevetti e requisiti
  - Analisi di dati utilizzati in ambito forense
  - ….

Text-to-Knowledge (T2K)

- Ed ora vediamo T2K all'*opera*:
  - *www.italianlp.it/demo/t2k-text-to-knowledge*