Università degli Studi di Modena e Reggio Emilia

Dipartimento di Ingegneria "Enzo Ferrari"

Corso di Laurea Magistrale in Ingegneria Informatica

Ibridazione di Machine Learning e Semantica per la Named-Entity Recognition

Candidato:

Antonio Circiello

Relatore:

Prof.ssa Sonia Bergamaschi

Correlatore:

Dott. Andrea Cappelli

Anno Accademico 2015/2016

Agenda

- Named-Entity Recognition
- Approccio semantico
- Approccio machine learning
- Approccio ibrido
- Conclusioni

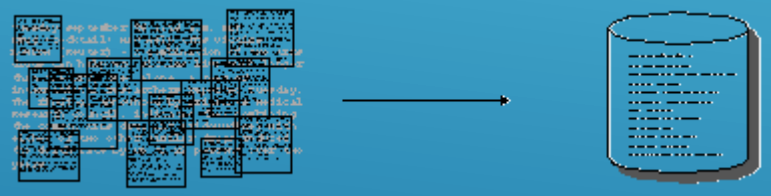
Named-Entity Recognition

Progetto di tesi

- Progetto di tesi svolto presso Expert System Spa
- Sistema di **Named-Entity Recognition** basato su **machine learning** e **semantica**
- Ibridazione di Conditional Random Fields e Disambiguazione
- Valutazione e test del sistema

Information Extraction

- Estrarre, in modo automatico, informazione strutturata a partire da documenti non strutturati o semi-strutturati
- **Information Extraction** include: Named-Entity Recognition, Coreference resolution, Relationshiop extraction
- NER: riconoscimento di nomi di entità conosciute



Named-Entity

- **Named-Entity** sono entità indicate con uno o più **rigid designator**
- Un designator è rigido se si riferisce alla stessa cosa in ogni possibile mondo

Il cittadino Mario Rossi vive a Roma

Il cittadino $[\text{Mario Rossi}]_{\text{PER}}$ vive a $[\text{Roma}]_{\text{LOC}}$

Definizione del problema

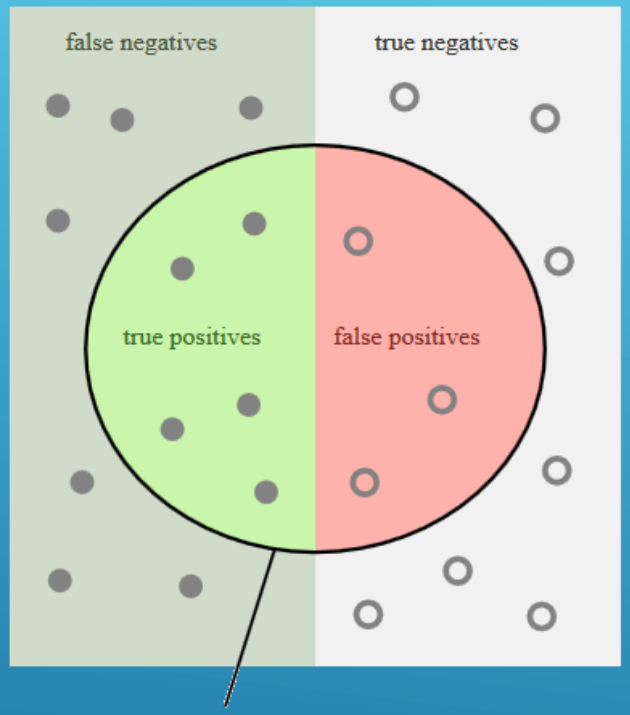
- Named-entity Recognition è spesso suddivisa in due distinti problemi: **rilevamento** dei nomi e **classificazione** in base al tipo di entità a cui si riferiscono

[Mario]$_{PER}$ [Rossi]$_{PER}$ 

[Mario Rossi]$_{LOC}$ 

[Mario Rossi]$_{PER}$ 

Elementi rilevanti



Elementi selezionati

- **Precision**: quanti elementi selezionati sono rilevanti?

- **Recall**: quanti elementi rilevanti sono stati selezionati?

- **F1 score**: media armonica di precision e recall

Parametri di valutazione

Approcci

Approccio semantico

- Rappresentazione della **conoscenza**
- Prestazioni più elevate
- Molto lavoro da parte di linguisti computazionali esperti

Approccio machine learning

- Focus su un metodo supervisionato
- Nessun accesso a conoscenza del mondo
- Grande quantità di **dati di training** annotati manualmente

Approccio semantico

- Individuare quale **significato** di una parola è usato in una particolare frase, quando la parola ha diversi significati.

Il calcio è un importante componente di una dieta equilibrata

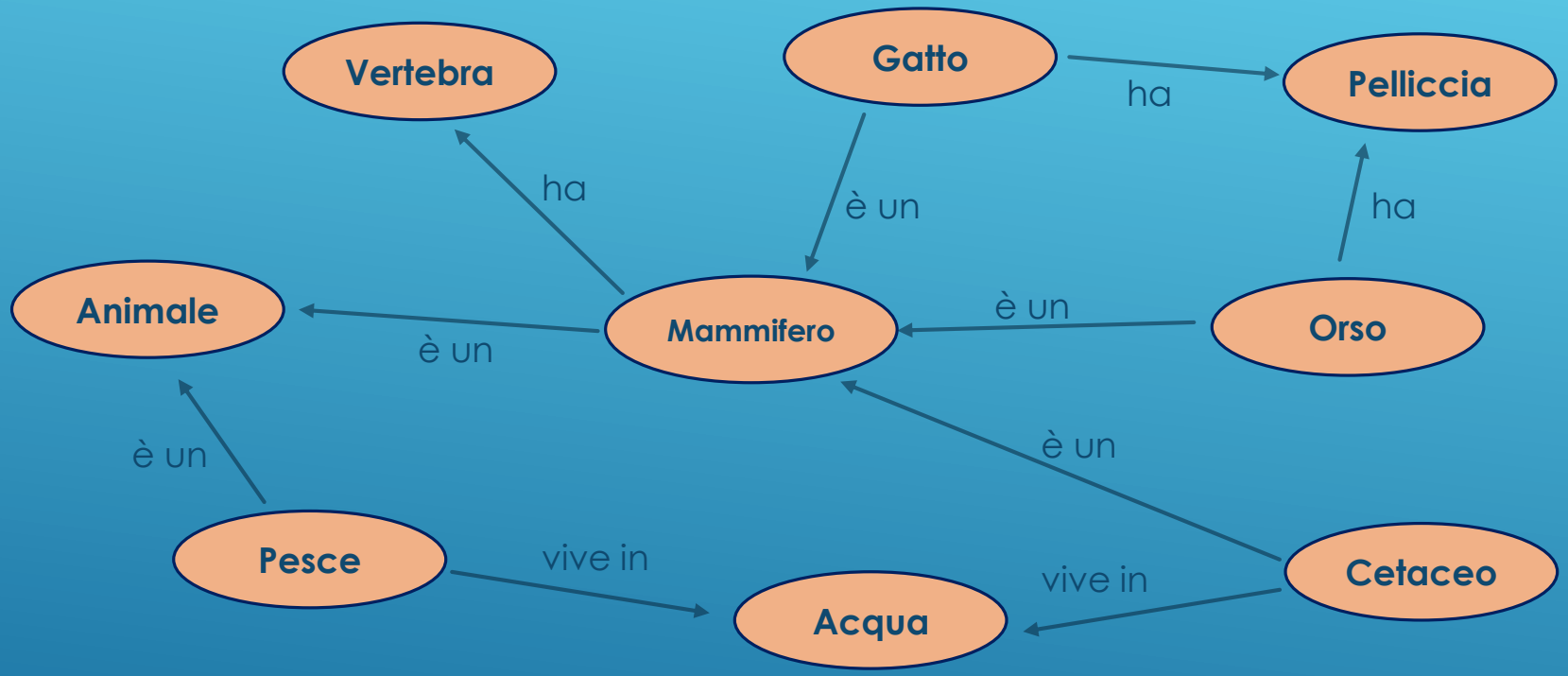
Calcio – elemento chimico

Il calcio è il mio sport preferito

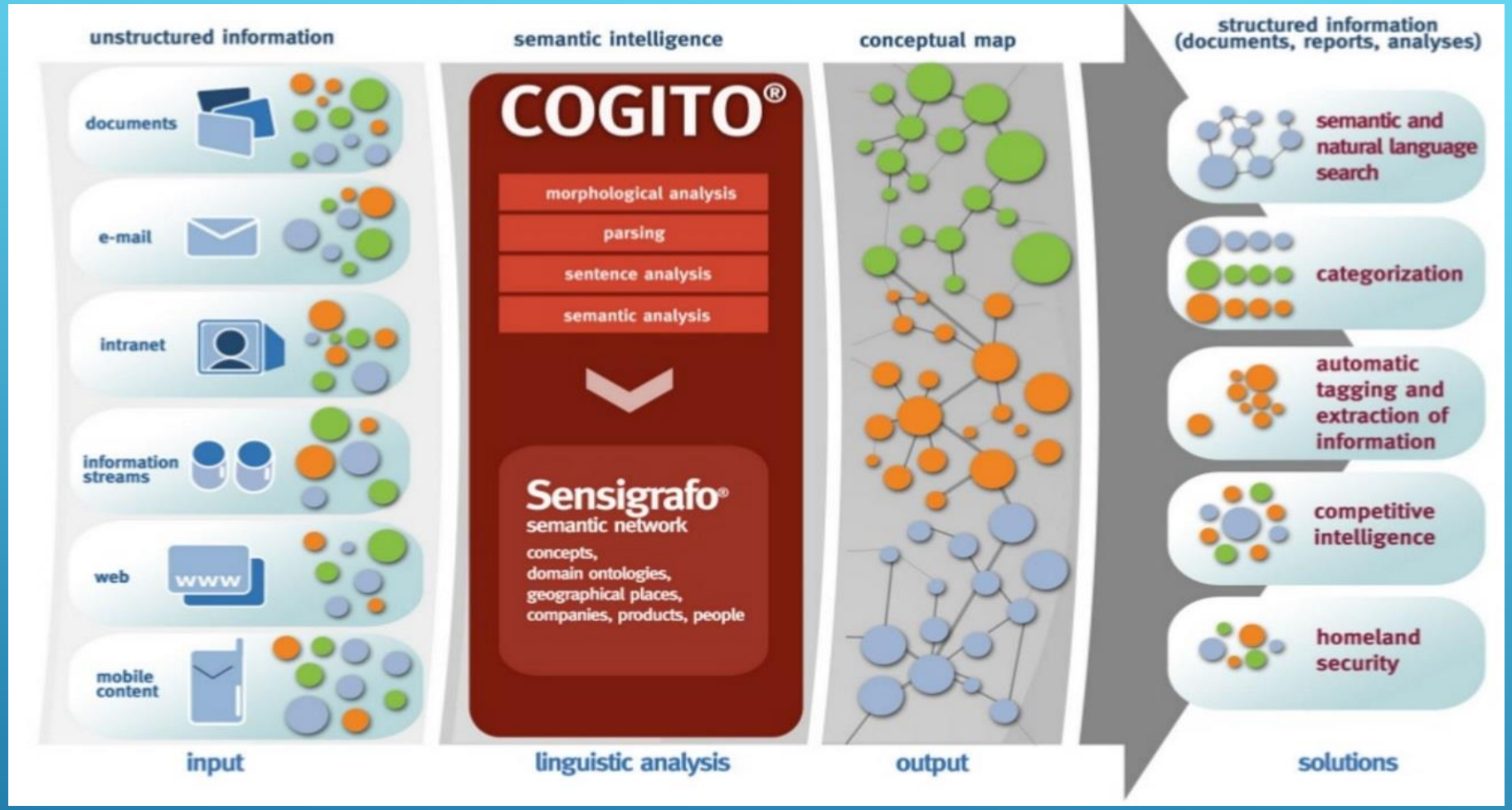
Calcio – sport

Word-Sense Disambiguation

- Una **rete semantica** è una rete che rappresenta delle **relazioni semantiche** tra dei concetti



Rete semantica

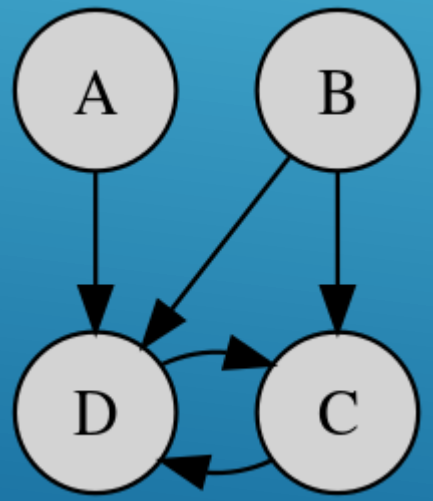


Tecnologia Expert System

Approccio machine learning

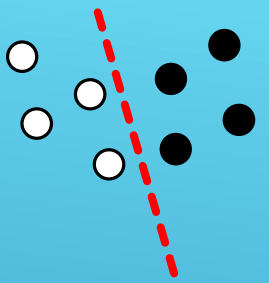
Graphical models

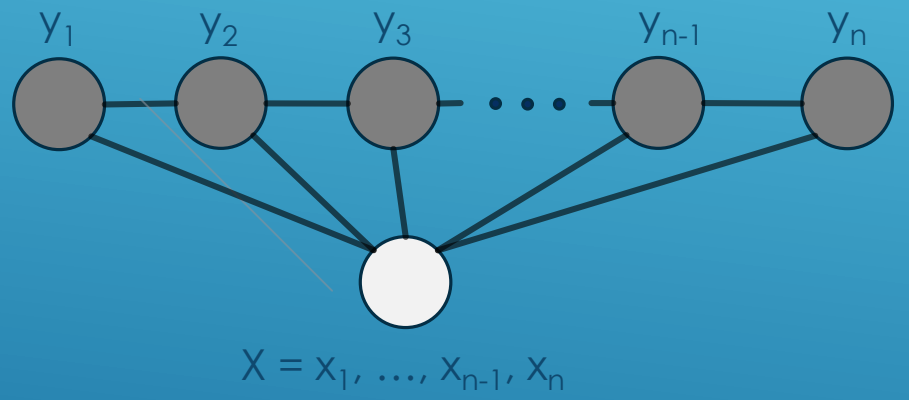
- Per applicazioni di **elaborazione del linguaggio naturale** è fondamentale essere in grado di predire più variabili dipendenti tra di loro
- Predire un vettore Y di variabili aleatorie dato un vettore di **osservazioni** X
- Un **graphical model** è un modello probabilistico per il quale un grafo esprime la struttura delle dipendenze tra le variabili aleatorie

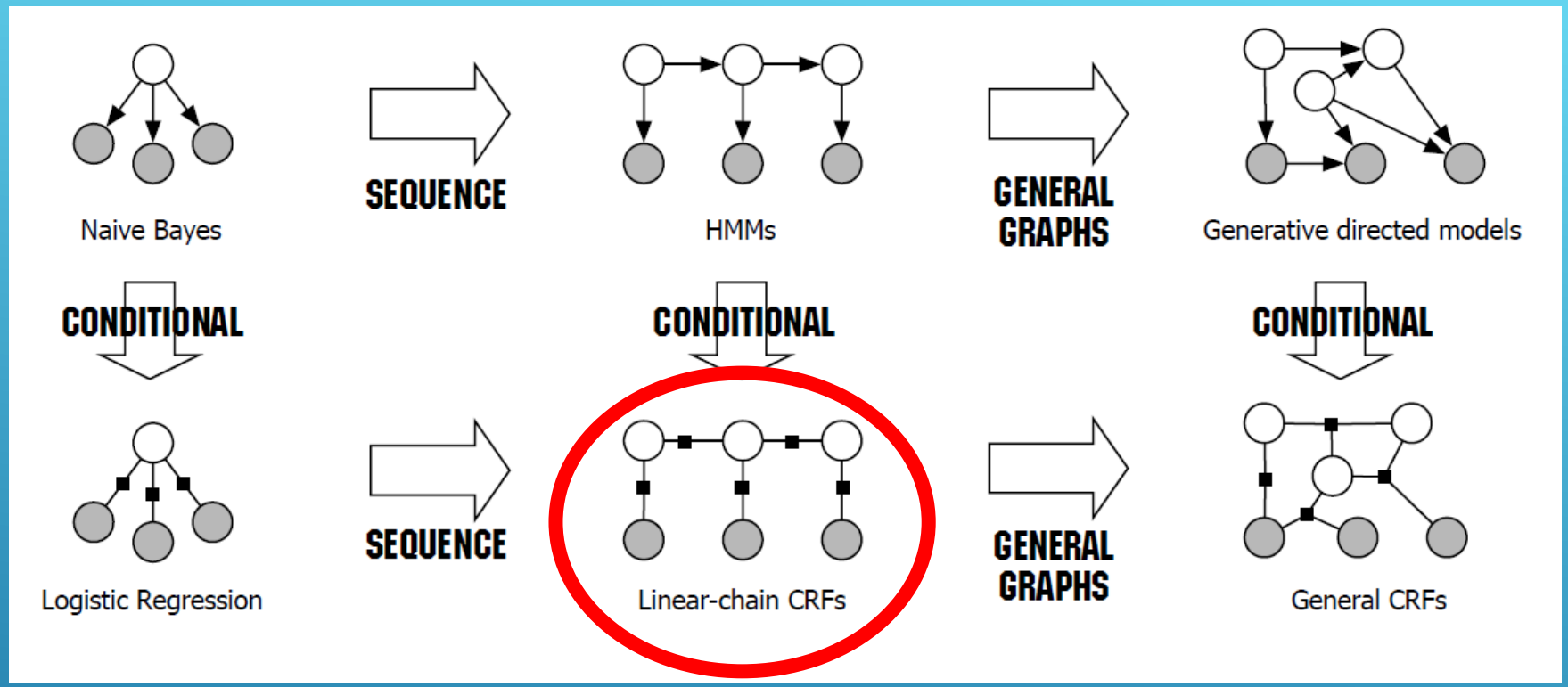


Conditional Random Fields

- Graphical model discriminativo, si ricava $P(Y|X)$ dai dati di training
- Mentre un ordinario classificatore produce una label per un singolo campione senza considerare i campioni "vicini", un CRF può considerare il **contesto**
- I linear-chain CRF, predicono **sequenze di label** per sequenze di campioni di input







Basato su feature functions

Conditional Random Fields

Feature functions

- Una **feature function** è una funzione che prende come input una frase, la posizione i di una parola nella frase, la label y_i della parola corrente, la label y_{i-1} della parola precedente e dà in output un **numero reale**
- Ad esempio, una feature function potrebbe misurare quanto una parola corrente dovrebbe essere taggata come LOC sapendo che la parola precedente era "Repubblica"

$$score(Y|X) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(X, i, y_i, y_{i-1})$$

Mario Rossi vive a Roma

PER PER O O LOC → $score(Y'|X)$

Mario Rossi vive a Roma

PER PER O O PER → $score(Y''|X)$

Mario Rossi vive a Roma

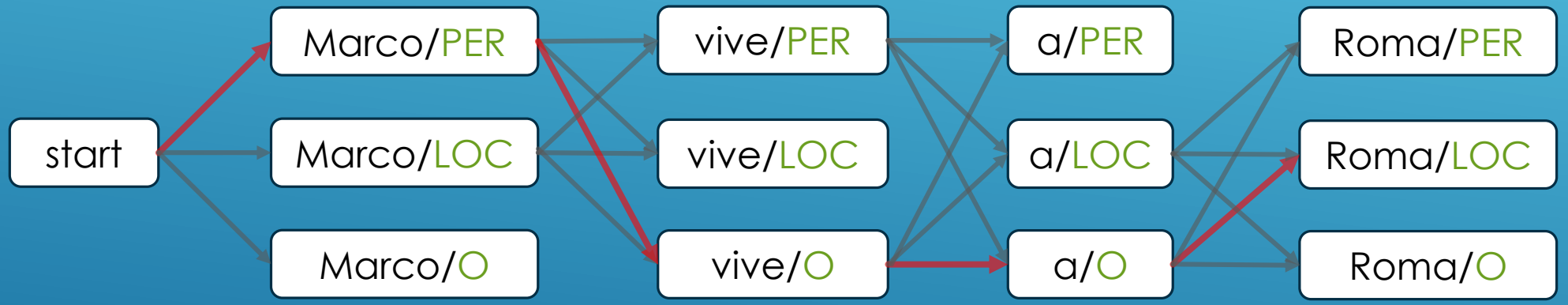
PER PER LOC O LOC → $score(Y'''|X)$

⟩ Y'

Labeling

Labeling

- Ci sono k^m label possibili per un tag set di dimensione k e una frase di lunghezza m token
- Un modo migliore per trovare la sequenza di label ottima è sfruttare un algoritmo di programmazione dinamica, quale l'**algoritmo di Viterbi**



Approccio ibrido

Sistema NER

- Python
- **Wapiti** (training, labeling, dump)
- **Sensigrafo** e **Disambiguatore**
- Vari tipi di entità
- Preparazione dei dati
- Calcolo statistiche

Dati di training

- Lingua inglese
- Tipi di entità: PER, LOC, ORG, MISC, O
- Convenzione **BIO**

A	O
U.S.	B-LOC
F-14	B-MISC
military	O
plane	O
landed	O
at	O
Ben	B-LOC
Gurion	I-LOC
airport	O

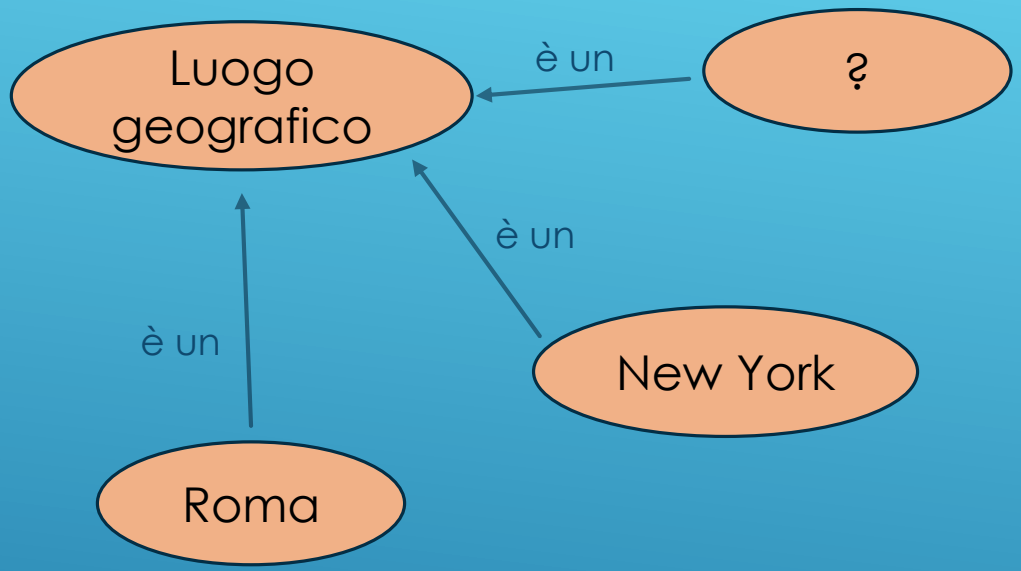
Feature non semantiche

- Token
- Tipo grammaticale (*)
- Forma base (*)
- Suffissi e prefissi
- Soddisfacimento o meno di alcune condizioni
- Pattern di composizione della parola

→ Feature dipendenti dal dominio

John24 → Aaaadd → Aad

(*) Non sono informazioni semantiche ma vengono comunque ottenute per mezzo del Disambiguatore



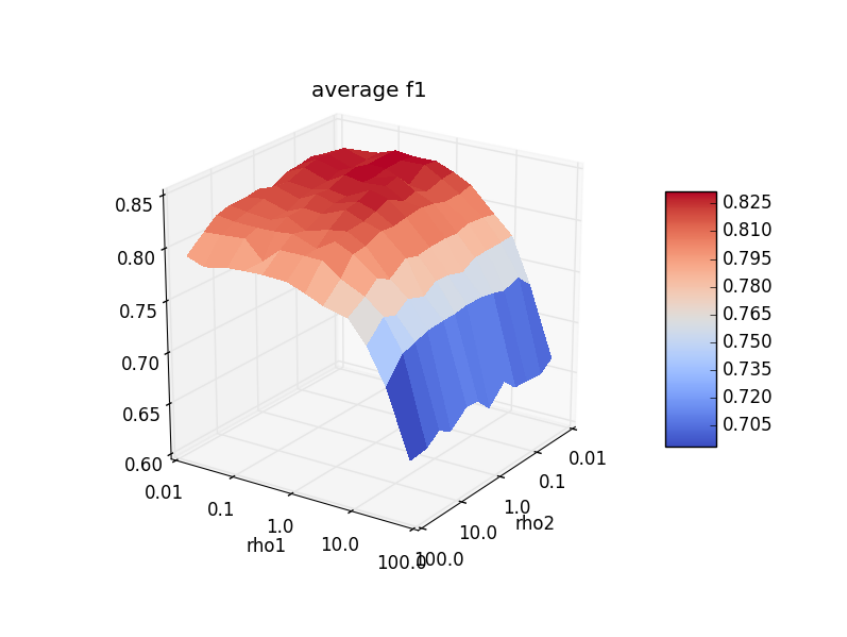
- **Ancestor** nel Sensigrafo
- Verbo di cui il token è soggetto/oggetto
- Tipo/categoria del documento

Feature semantiche

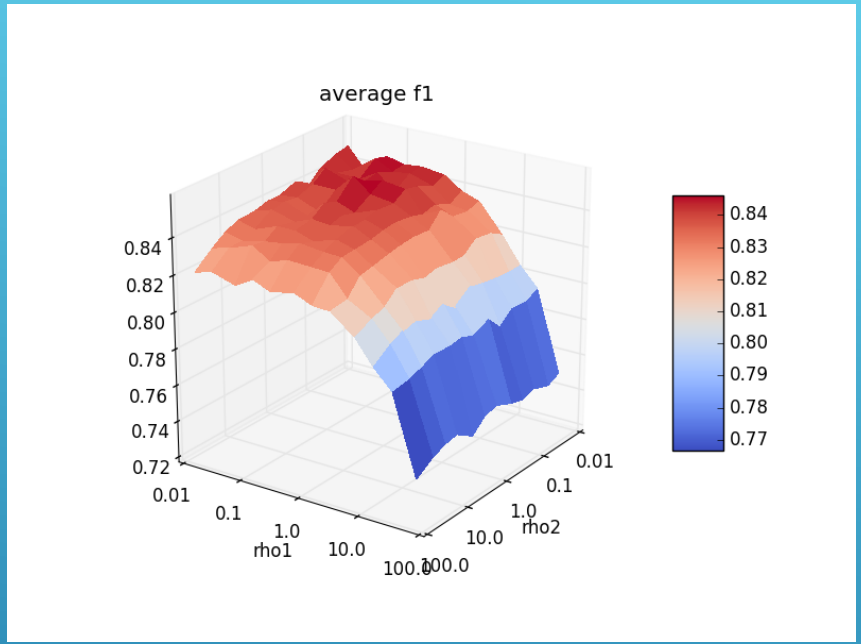
1. Confronto tra sistema NER con e senza informazioni semantiche
2. Confronto tra sistema NER con e senza informazioni semantiche al variare delle dimensioni del corpus
3. Miglioramento dei modelli mediante analisi delle feature
4. Stress testing
5. Sviluppo di regole di estrazione

Esperimenti fatti

Modello senza informazioni semantiche



Modello con informazioni semantiche



Confronto tra modelli con e senza semantica

Modello senza informazioni semantiche

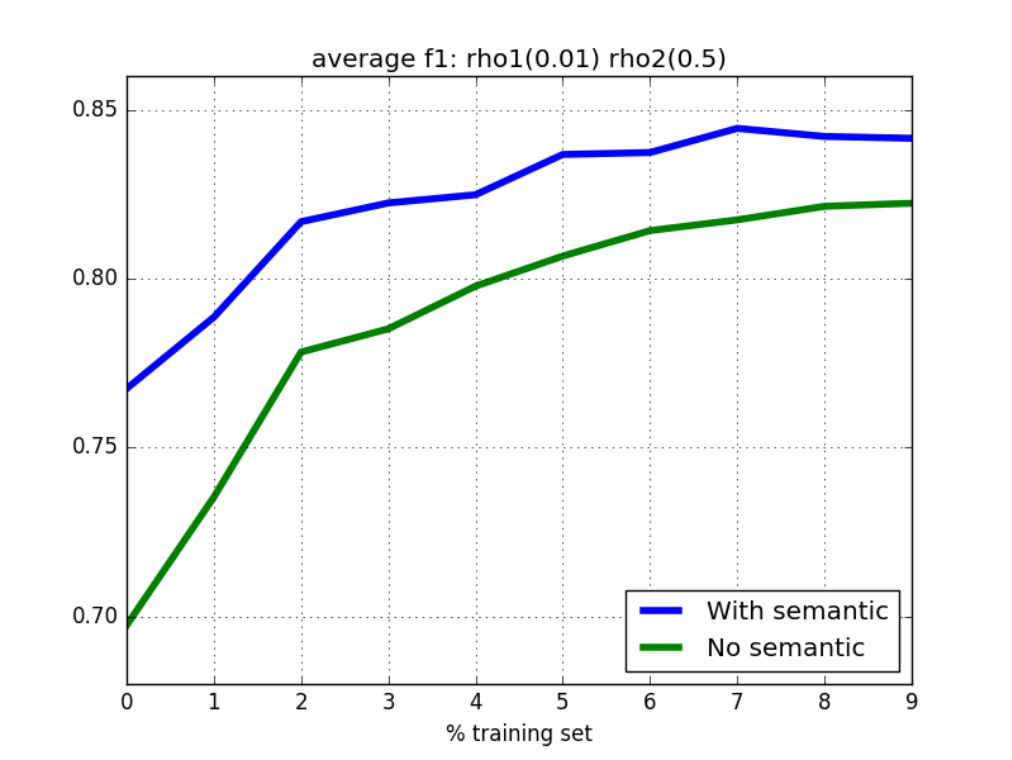
$\rho_1 = 0.2$ e $\rho_2 = 0.1$			
	PRECISION	RECALL	F1-SCORE
LOC	0.8786	0.8775	0.8780
PER	0.8780	0.9092	0.8933
ORG	0.8373	0.7379	0.7845
MISC	0.8090	0.7507	0.7787
Average	0.8507	0.8188	0.8336
Overall	0.8590	0.8287	0.8441

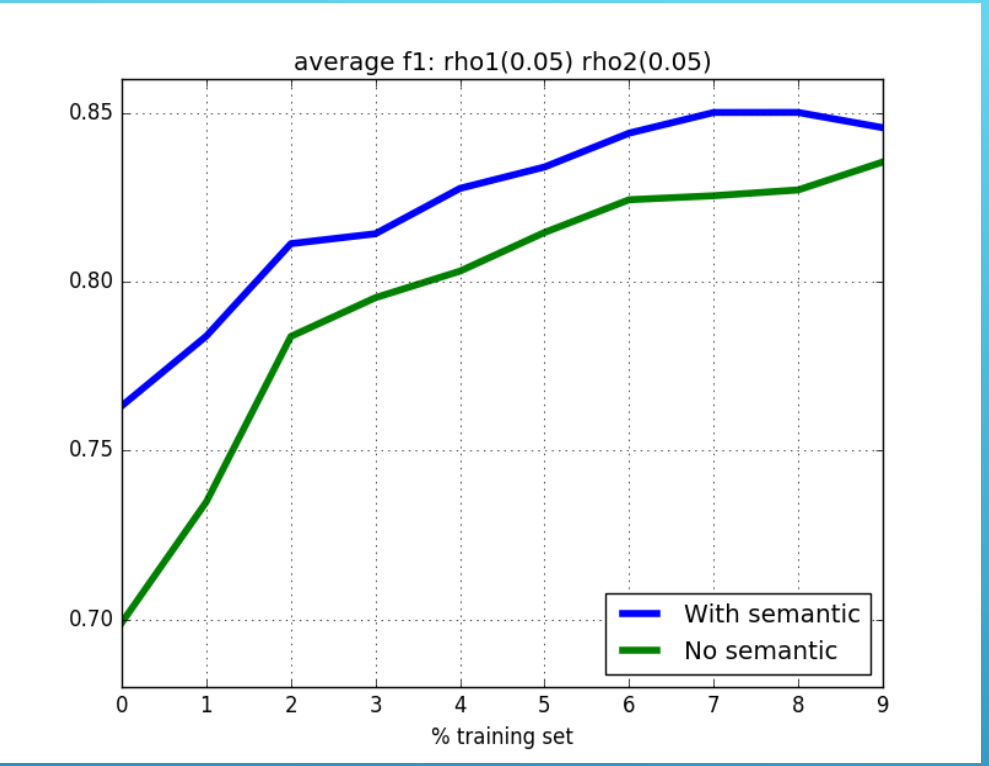
Modello con informazioni semantiche

$\rho_1 = 0.2$ e $\rho_2 = 0.2$			
	PRECISION	RECALL	F1-SCORE
LOC	0.9004	0.8896	0.8950
PER	0.9033	0.9297	0.9163
ORG	0.8250	0.7813	0.8025
MISC	0.8228	0.7564	0.7882
Average	0.8629	0.8392	0.8505
Overall	0.8707	0.8526	0.8616

Differenza ≈ 2 %

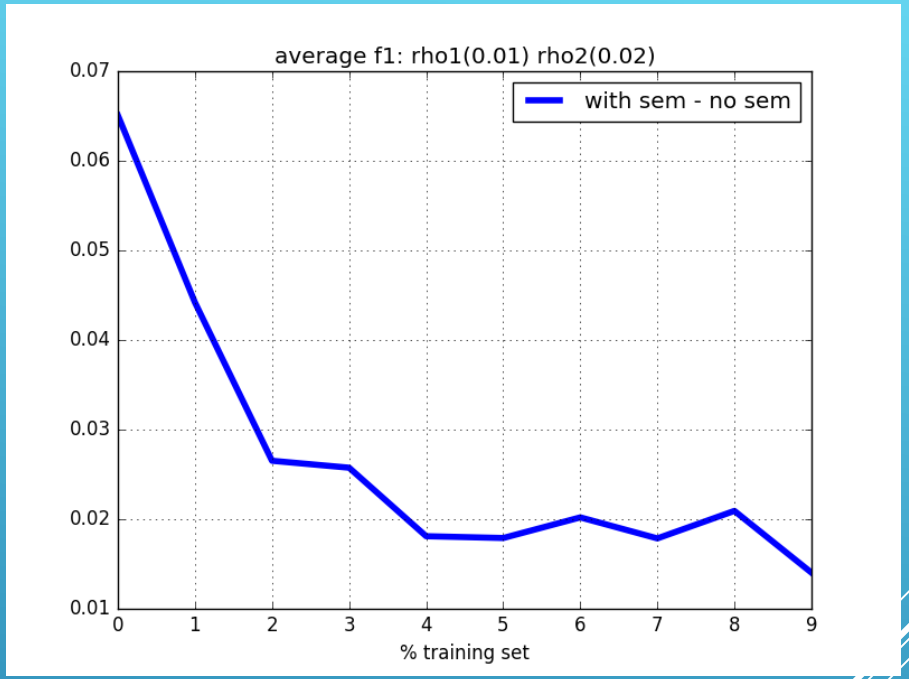
Confronto tra modelli con e senza semantica





Confronto al variare della dimensione del corpus

Dimensione del corpus rispetto al corpus intero	AVERAGE F1-SCORE (senza semantica)	AVERAGE F1-SCORE (con semantica)	DIFFERENZA (%)
10 %	0.7069	0.7697	6.27 %
20 %	0.7515	0.7923	4.07 %
30 %	0.7874	0.8183	3.09 %
40 %	0.7975	0.8272	2.96 %
50 %	0.8108	0.8312	2.03 %
60 %	0.8207	0.8387	1.79 %
70 %	0.8254	0.8438	1.84 %
80 %	0.8317	0.8508	1.91 %
90 %	0.8336	0.8529	1.93 %
100 %	0.8387	0.8553	1.65 %

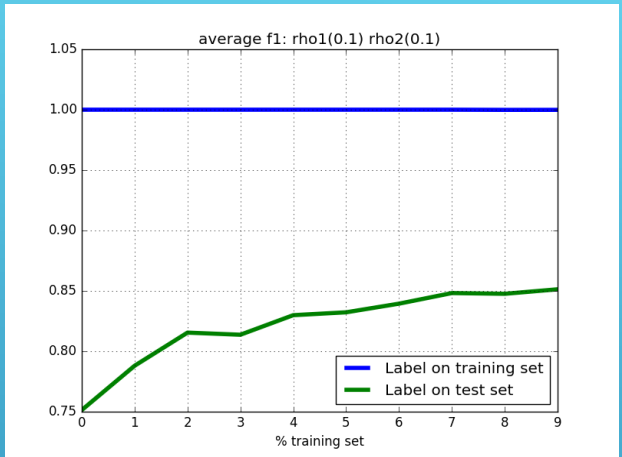


Confronto al variare della dimensione del corpus

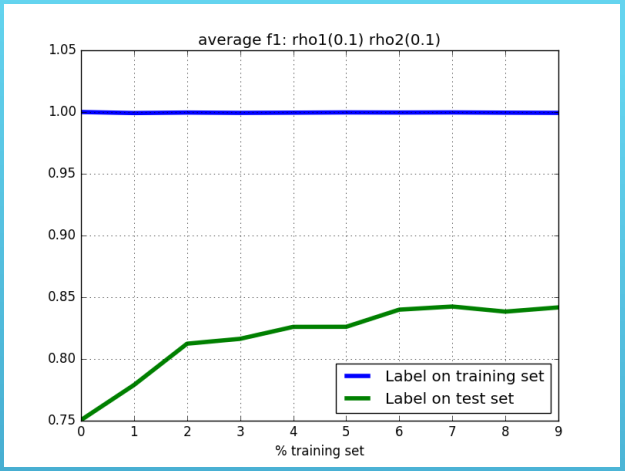
Prime 20 feature (su 125 totali), ordinate in base al valore assoluto del peso medio

1. {'name': 'u:grammar_type(0)', 'ave_weight': 0.7127867368421054}
2. {'name': 'u:anycharisdigit(0)', 'ave_weight': 0.5261616666666666}
3. {'name': 'u:anycharupper(0)', 'ave_weight': 0.4841533333333333}
4. {'name': 'u:firstcharupper(0)', 'ave_weight': 0.4355972222222222}
5. {'name': 'u:isdigit(0)', 'ave_weight': 0.2901222222222223}
6. {'name': 'u:word_composition2(0)', 'ave_weight': 0.28903201492537295}
7. {'name': 'u:startofsentence(0)', 'ave_weight': 0.20896444444444442}
8. {'name': 'u:dashintoken(-1)', 'ave_weight': 0.20493777777777777}
9. {'name': 'u:word_compisotion1(0)', 'ave_weight': 0.19725171091445465}
10. {'name': 'u:anycharisalpha(0)', 'ave_weight': 0.19123222222222222}
11. {'name': 'u:ancestor1(0)', 'ave_weight': 0.1860191489361707}
12. {'name': 'u:domain(0)', 'ave_weight': 0.18539940000000013}
13. {'name': 'u:isalpha(0)', 'ave_weight': 0.17757888888888887}
14. {'name': 'u:onlyfirstcharupper(0)', 'ave_weight': 0.17650833333333335}
15. {'name': 'u:ancestor2(0)', 'ave_weight': 0.1638238918205809}
16. {'name': 'u:ancestor4(0)', 'ave_weight': 0.14496880782918165}
17. {'name': 'u:grammar_type(-1)', 'ave_weight': 0.14416326923076922}
18. {'name': 'u:ancestor3(0)', 'ave_weight': 0.14161505882352973}
19. {'name': 'u:anycharisdigit(-1)', 'ave_weight': 0.14120148148148143}
20. {'name': 'u:grammar_type(+1)', 'ave_weight': 0.12482905660377364}

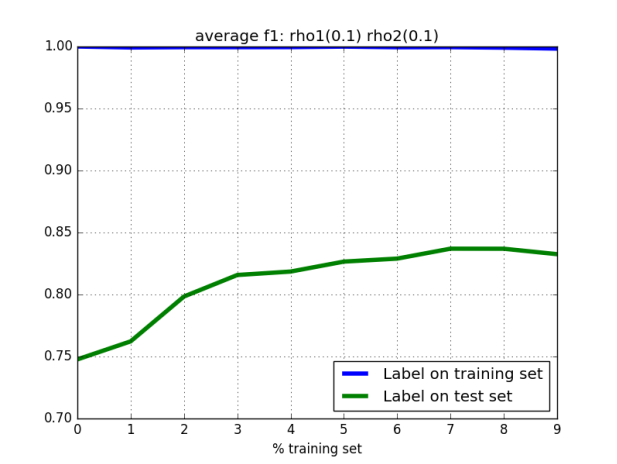
Dump di un modello



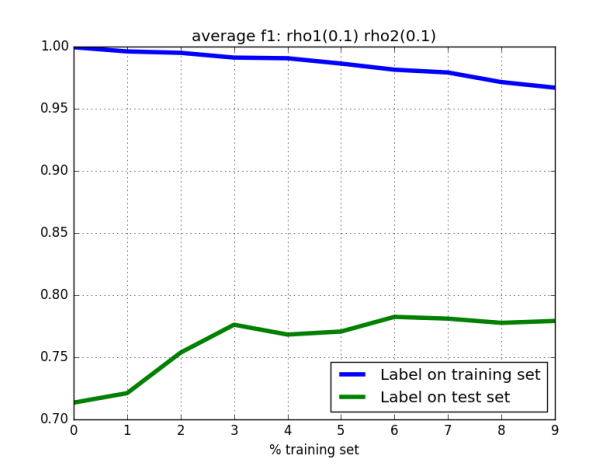
1ª iterazione



5ª iterazione



6ª iterazione

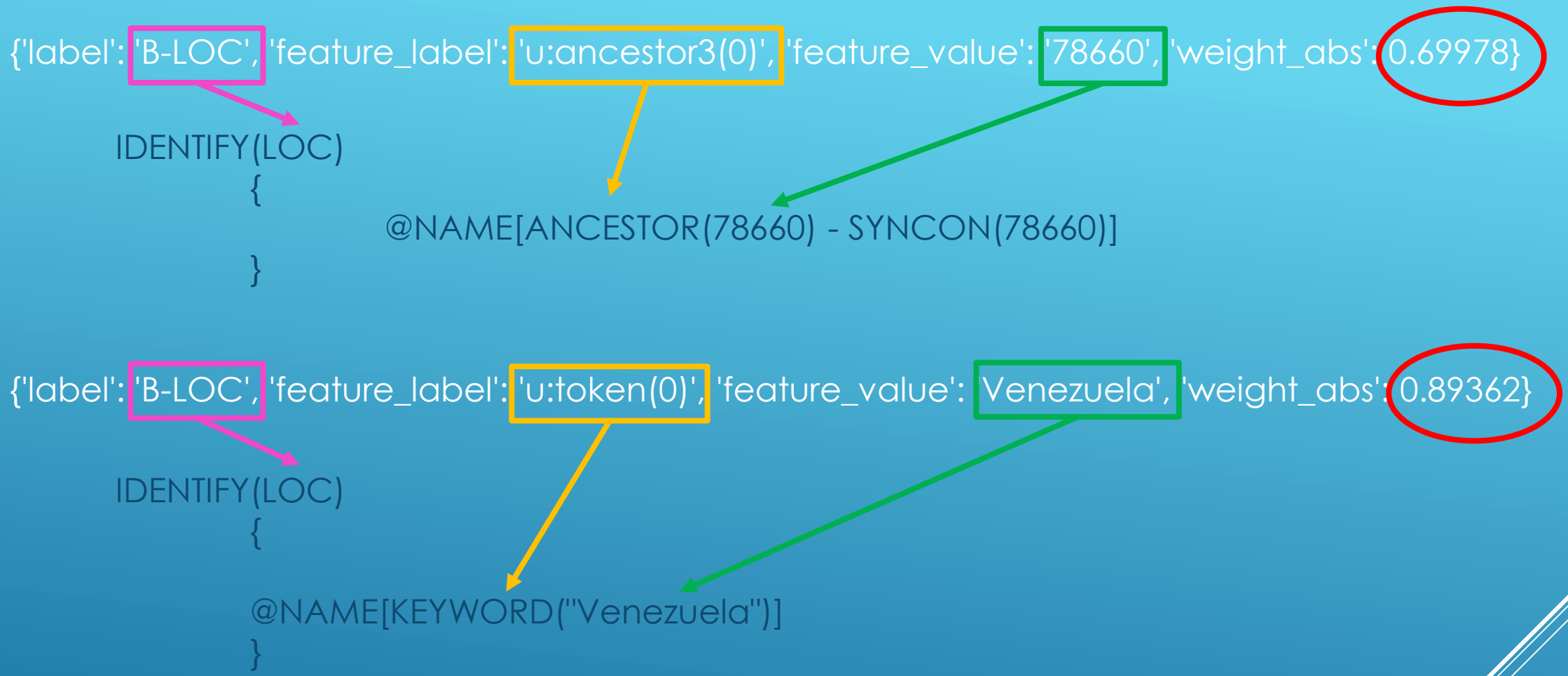


8ª iterazione

Analisi delle feature

Estrazione

- Progetti di text mining basati sull'analisi semantica dei testi
- L'**estrazione** è un processo in cui i dati sono estratti da varie fonti, trasformati in modo da essere adeguati alle necessità e resi disponibili per ulteriori processamenti
- I dati sono identificati ed estratti da un documento per mezzo di **regole linguistiche**



Regole di estrazione

Conclusioni

- Prove su corpus spagnolo
- Prove su corpus con entità di tipo diverso
- Miglioramento delle features
- Integrazione

Conclusioni

Grazie per l'attenzione!