LISA - Light Italian Semantic Analyzer

Un tagger semantico *leggero* per l'italiano

Ludovica Pannitto
Relatore: Prof. Alessandro Lenci
Correlatore: Dott. Felice Dell'Orletta

7 luglio 2016

Corso di Laurea in Informatica Umanistica - Università di Pisa

Obiettivo

Il nostro obiettivo è di costruire un modello che sia in grado di distinguere il senso che assume una parola a seconda del contesto in cui si trova. Ad esempio, noi percepiamo diversi i sensi della parola *consiglio* nei seguenti casi:

(1) Potresti darmi un *consiglio* su quale vestito indossare?

(2) Il *consiglio* ha deliberato il nuovo regolamento per le lauree

(3) La giunta si è riunita nella sala del *consiglio*

(4) Ci sono state manifestazioni durante lo svolgimento del *consiglio*

Contenuti

1. SuperSense Tagging

2. Light Semantic Tagging

3. Creazione del modello

4. Conclusioni e Sviluppi Futuri

SuperSense Tagging

SST

- Il SuperSense Tagging (SST) consiste nell'annotare ogni entità in un contesto con la categoria giusta in riferimento ad una certa tassonomia
- A metà strada tra un task di Word Sense Disambiguation (WSD) e un task di Named Entity Recognition (NER)
- Vocazione generale, prende in considerazione tutti i nomi, pochi sensi e semplici da identificare
- Termine *supersenso* coniato in Ciaramita e Johnson 2003
- Derivato dalle classi lessicografiche di WordNet

Stato dell'Arte

Tabella 1: Stato dell'Arte

	lingua	N	V	ADJ	ADV	Miglior risultato	Algoritmo
Ciaramita e Johnson 2003	eng	x				53.4% (accuracy)	Multiclass Averaged Perceptron
Curran 2005	eng	x				68% (accuracy)	Semantic Similarity
Ciaramita e Altun 2006	eng	x	x			77.18% (f-measure)	Hidden Markov Model
Picca et al. 2008	ita	x				62.9% (f-measure)	Hidden Markov Model
Attardi et al. 2010	ita	x	x	x	x	79.1% (f-measure)	Maximum Entropy
Evalita 2011 - UniPi	ita	x	x	x	x	78.27% (f-measure)	Maximum Entropy
Evalita 2011 - UniBa	ita	x	x	x	x	75.34% (f-measure)	Support Vector Machine

EVALITA 2011

- Quattro categorie: Nomi, Verbi, Aggettivi e Avverbi
- Tagset derivato dalle classi lessicografiche di WordNet
- Due subtask: *open subtask* e *closed subtask*
- Corpus ISST-SST, con revisione manuale
- Partecipano due team (UniPi e UniBa) al closed subtask, solo UniBa all'open subtask

Light Semantic Tagging

Il nostro task

- Ci restringiamo alla sola categoria dei sostantivi
- Il corpus è stato annotato automaticamente, dunque senza revisione manuale a differenza della risorsa fornita per EVALITA 2011, fino al livello di parsing sintattico a dipendenze, aggiungendo dunque un livello di analisi
- L'ontologia di riferimento è diversa: il task di SuperSense Tagging si è consolidato come task di classificazione rispetto ai supersensi derivati da WordNet. Nel nostro task le categorie semantiche sono state sviluppate a partire dalla Light Semantic Ontology

Il task è stato affrontato come un task supervisionato di classificazione multiclasse, per la costruzione del modello è stata utilizzata una Support Vector Machine con kernel lineare

Light Semantic Ontology

- Necessità di una classificazione minimale e linguisticamente plausibile
- Problemanticità di alcuni supersensi derivati da WordNet

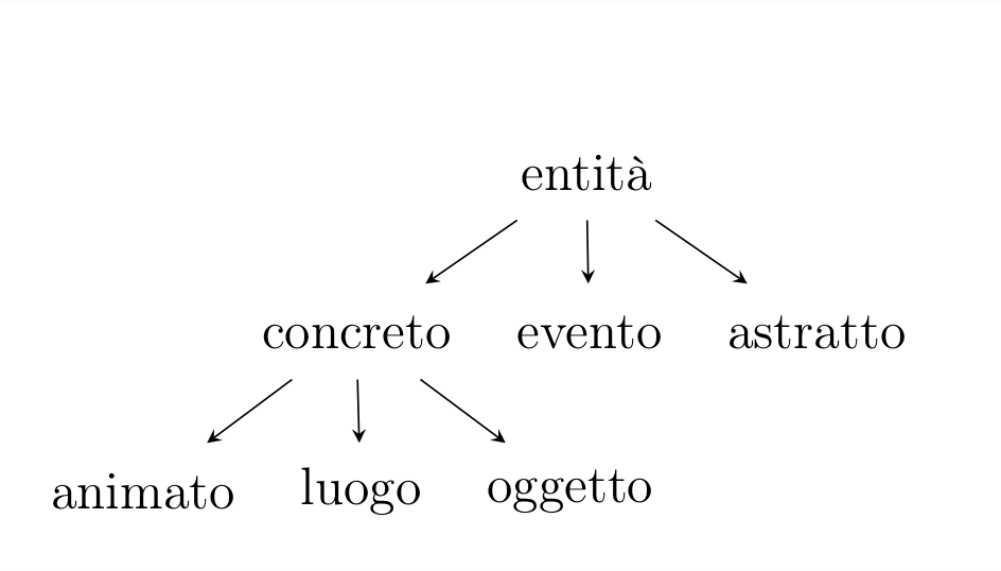

Figura 1: Schema Light Semantic Ontology

Supersensi

Tabella 2: Supersensi

person	cognition	attribute	quantity	plant
communication	possession	object	motive	relation
artifact	location	process	animal	
act	substance	Tops	body	
group	state	phenomenon	feeling	
food	time	event	shape	

Tagset

O(ther)	sostantivi privi di contenuto semantico perché utilizzati in costruzioni funzionali, o errori di PoS Tagging
Animate	sostantivi utilizzati in contesti agentivi
Location	sostantivi che esprimono collocazioni spaziali relativamente fisse e indicano qualcosa in cui ha luogo un evento
Object	sostantivi che indicano oggetti concreti (ad esempio artefatti) o sostanze, in generale che denotano entità percepibili attraverso i sensi
Event	sostantivi che esprimono entità che accadono nel tempo
Abstract	entità che si riferiscono a concetti astratti, come sentimenti, ideali, ...

- I tag sono stati utilizzati in formato IOB
- Sono state considerate MultiWord Expression solo le sequenze rigidamente non composizionali.
- I gruppi di entità sono stati etichettati come le entità di cui è formato il gruppo.
- Entità che specificano un sintagma nominale che ne descrive il tipo, tendono ad ereditare il tipo della testa dell'NP padre.

487 documenti divisi in frasi, circa 314 mila token, di cui quasi 86 mila sostantivi.

Tabella 3: Composizione Sostantivi

PoS	Descrizione	Token
S	Sostantivi	64303
SP	Nomi Propri	20759
SW	Sostantivi Stranieri	643
SA	Abbreviazioni	258

Tabella 4: Distribuzione istanze nelle classi del tagset

Classe	Token	Lemmi
ABSTRACT	34785	5219
ANIMATE	23662	6255
EVENT	9362	2087
LOCATION	8029	2203
OBJECT	5248	1847
O	4877	1544

Annotazione

L'affidabilità dell'annotazione è stata testata calcolando l'intercoder agreement tra gli annotatori. I risultati mostrano un accordo soddisfacente: simple agreement - Fleiss' κ: 76.1%, con $p < 0.001$

$$\kappa = \frac{A_o - A_e}{1 - A_e} \tag{1}$$

Alcune classi risultano meno facilmente identificabili di altre, come riportato in tabella:

	Fleiss'-κ	**p-value**
ABSTRACT	0.716	0.000
ANIMATE	0.902	0.000
EVENT	0.655	0.000
LOCATION	0.833	0.000
OBJECT	0.654	0.000

Tabella 5: Dettaglio agreement per classe

Esempi

Casi di **omonimia e polisemia**:

(5) a. Il *caccia*$_{\text{OBJECT}}$ è troppo veloce e il radar troppo poco potente; [...].

b. La ragazza è stata rilasciata dopo cinque ore e si è ripresentata a casa sconvolta, nel pieno della *caccia*$_{\text{EVENT}}$ ai sequestratori.

(6) a. Nel 1993 _ cito sempre dal rapporto Anee _ sono stati venduti 2,7 milioni di *lettori*$_{\text{OBJECT}}$ di cd-rom negli Stati Uniti, 900 mila in Asia e 400 mila in Europa.

b. Ho visto subito forte l' Argentina, come sanno i miei *lettori*$_{\text{ANIMATE}}$.

Fenomeni ricorrenti - **alternanza luogo/animato**

(7) a. Lunga catena di vittime, *Lombardia*$_{\text{ANIMATE}}$ in lutto

b. Ho telefonato anche a istituti della *Lombardia*$_{\text{LOCATION}}$ e dell'Emilia Romagna

Fenomeni ricorrenti - **alternanza oggetto/astratto**

(8) a. Il *libro*$_{\text{OBJECT}}$ era di cento pagine e costava sei lire.

b. La scelta del titolo d'un *libro*$_{\text{ABSTRACT}}$ è spesso motivo di angustie per l'autore.

Inoltre, **usi metaforici** e **sottospecificazione**:

(9) Il Presidente ha preso *la palla* al balzo

Creazione del modello

Fasi di Lavoro

- Individuazione delle knowledge sources e feature extraction
- Selezione del modello
- Ranking delle feature tramite Recursive Feature Elimination
- Valutazione del modello

Features

Features del Token Corrente Lemma, Part of Speech, Morfologia, WordShape, Lemma-PoS, Features dell'aspetto ortografico

Features Locali Pattern Locali di PoS, Pattern Locali per la forma ortografica, dipendenze sintattiche (Determinanti e modificatori numerali, modificatori aggettivali, Testa e Dipendenti), preferenze di selezione

Features Globali Preferenze di supersensi

Fonti: corpus di Repubblica[1], LexIt[2]

[1] `http://sslmit.unibo.it/repubblica`
[2] `http://lexit.fileli.unipi.it/`

Support Vector Machine

I modelli sono stati creati con una Support Vector Machine con Kernel Lineare e valutati sul training set tramite una 10-fold cross validation.

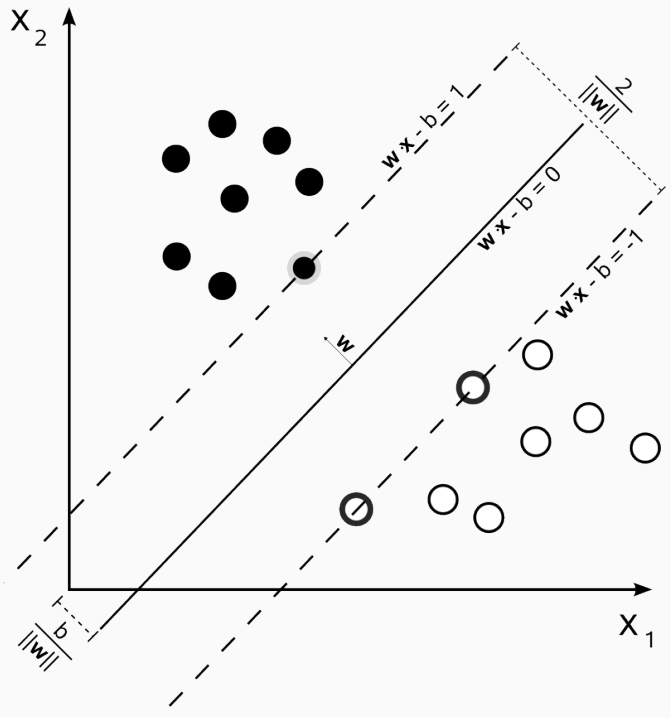

Figura 2: Massimizzazione del margine nella Support Vector Machine

Selezione del modello

- La versione base è costituita dalle features del modello presentato dal team dell'UniPi a EVALITA 2011, ad esclusione di lemmi e word forms
- Sono stati creati altri cinque modelli più uno, comprendente tutte le features

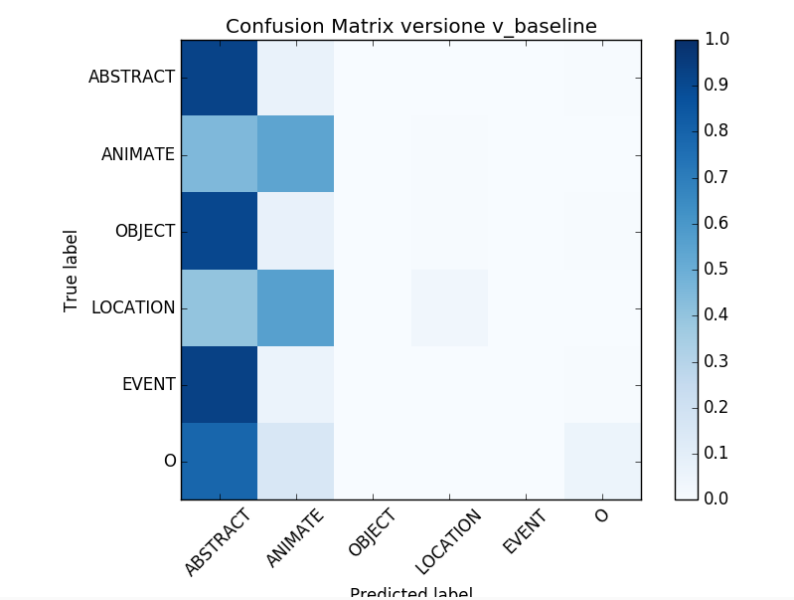

Figura 3: Nucleo base di features

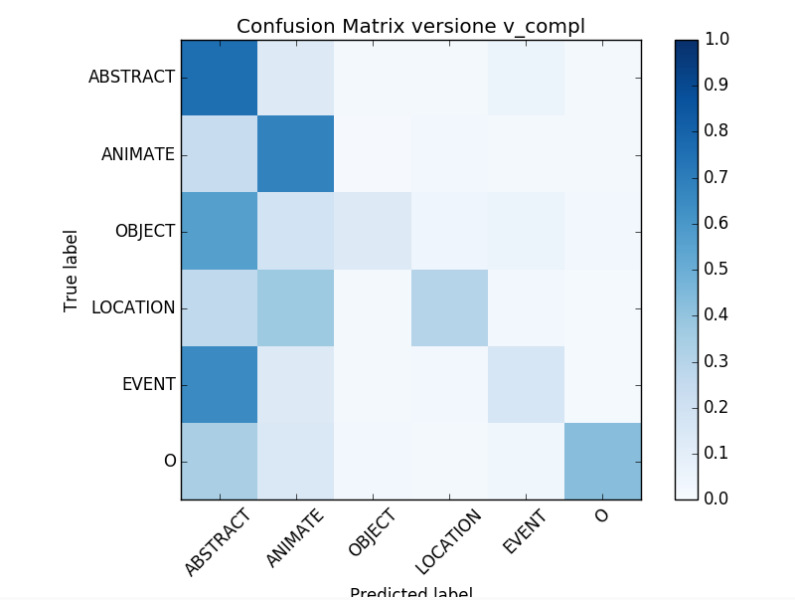

Figura 4: Versione completa

La classe dei nomi astratti sembra la più semplice da idenfiticare: ciò è sicuramente da attribuirsi alla maggior frequenza di questa classe rispetto alle altre

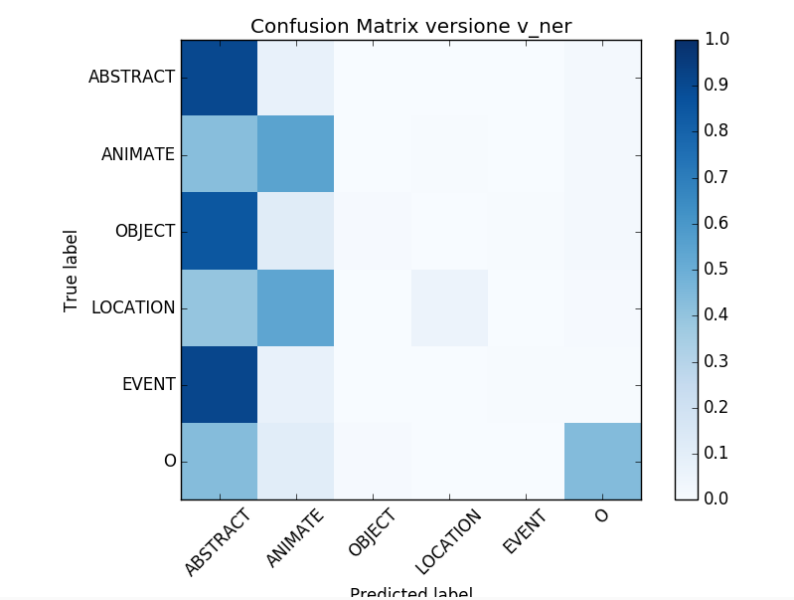

Figura 5: Nucleo base di features + ner

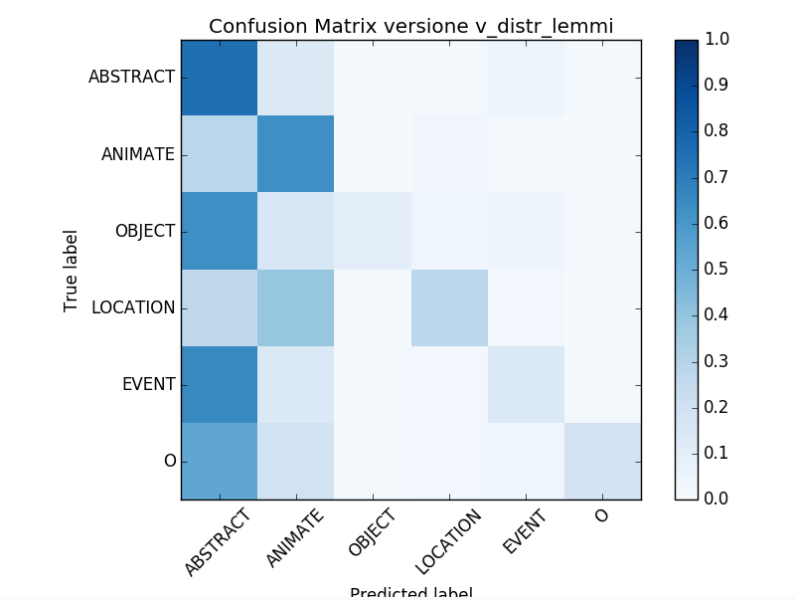

Figura 6: Nucleo base di features + informazione sintattica

Le feature del gruppo *ner* si mostrano più che valide per il riconoscimento della classe OTHER.

L'informazione sintattica permette di distinguere, seppur in minima percentuale, le classi a bassa frequenza (OBJECT, LOCATION, EVENT), dimostrandosi dunque valida, anche se non sufficiente, per l'obiettivo da noi considerato.

Valutazione dei risultati

Il modello completo di tutte le features estratte è stato valutato sulla porzione di test del corpus. Lo stesso è stato fatto per il modello *base*, che considriamo come la nostra baseline, e il modello *unipi*, creato a partire dall'assetto presentato a EVALITA dal team UniPi.

I risultati rivelano una maggiore stabilità dell'informazione non lessicale da noi selezionata rispetto all'insieme di features su cui il modello *unipi* si basa.

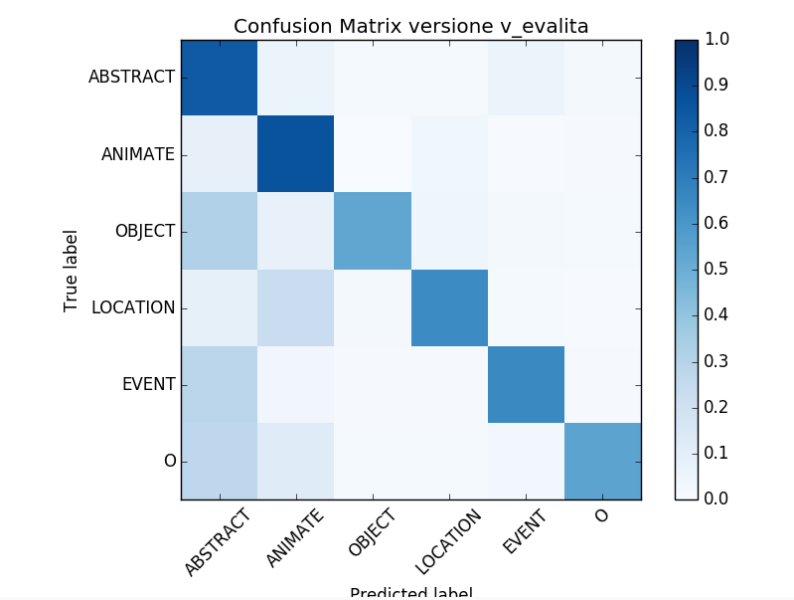

Figura 7: Assetto presentato dal team UniPi a Evalita 2011

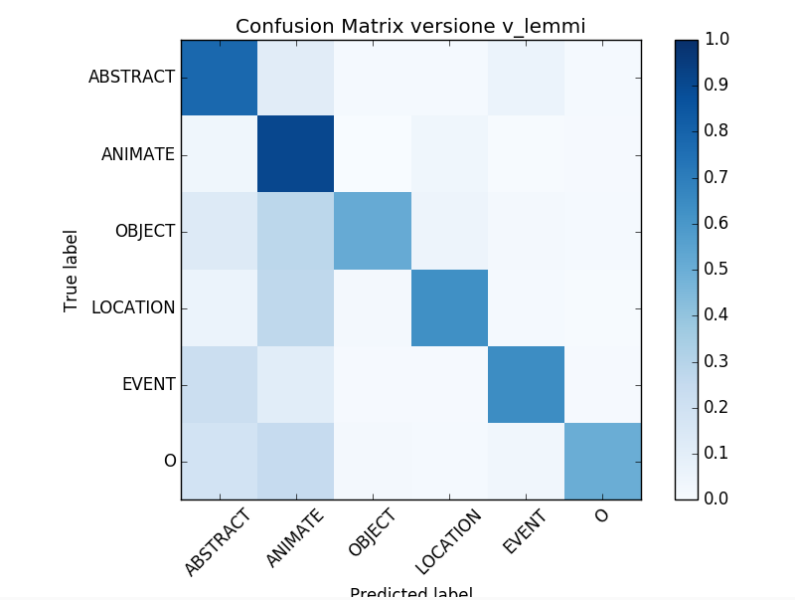

Figura 8: Classificazione ottenuta con i soli lemmi

L'informazione lessicale si conferma una valida fonte di disambiguazione, grazie anche ai livelli di *ambiguità* presenti nel corpus.

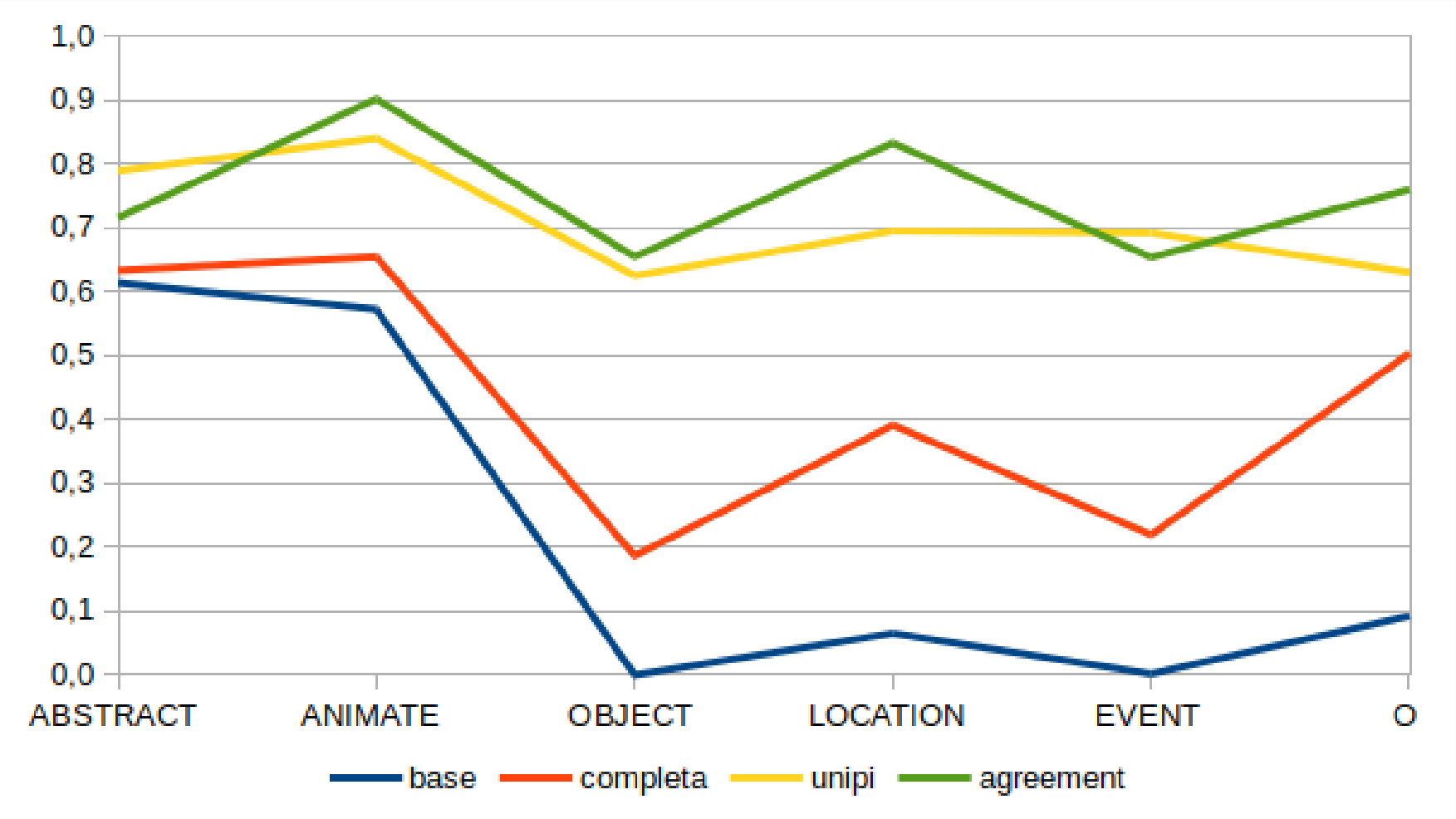

Figura 9: Confrontro tra *baseline*, versione *completa* e features del modello *unipi*

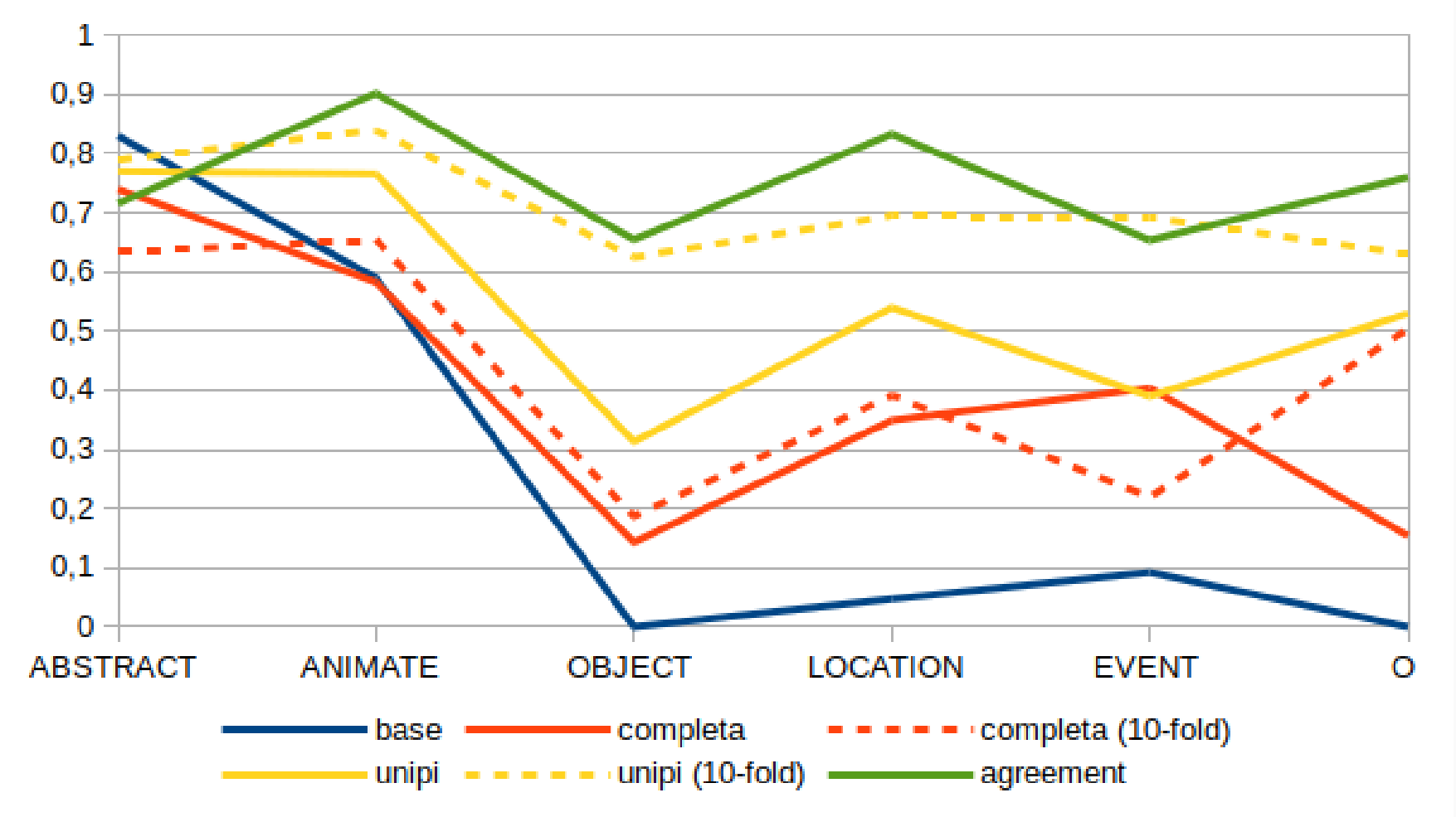

Figura 10: Valutazione del modello sul test set

Conclusioni e Sviluppi Futuri

Conclusioni

- Tra i modelli citati e il nostro le differenze sono moltissime: granularità dei tagset, qualità delle risorse usate, algoritmi, ontologia di riferimento
- L'informazione sintattica e semantica, come collocazioni o associazioni semantiche, forniscono apporto limitato. Questo (come suggerito in Agirre e Stevenson 2007) potrebbe essere dovuto alla qualità del parsing
- Bisogna considerare anche l'interazione tra le fonti di informazione e l'algoritmo scelto

Futuri Sviluppi

- Iperparametri della SVM
- Analisi più approfondita degli errori
- Miglioramenti nell'estrazione delle features (n-grammi, funzioni di smoothing, dizionari e filtri...)
- Esplorazione di campi che si sono dimostrati validi in letteratura: riconoscimento preliminare di MWE, uso di informazione globale, tecniche di modellazione come *word embeddings*

Grazie :)

Ambiguità

La distribuzione dei sostantivi nelle varie classi derivate dall'ontologia risulta molto sbilanciata

Tabella 6: Distribuzione istanze nelle classi del tagset

Classe	**Token**	**Lemmi**
ABSTRACT	34785	5219
ANIMATE	23662	6255
EVENT	9362	2087
LOCATION	8029	2203
OBJECT	5248	1847
O	4877	1544

Se consideriamo il livello di ambiguità di ogni lemma, scopriamo che i lemmi non ambigui rappresentano l'80.1% del totale, coprendo però solo il 38.9% dei token.

Tabella 7: Copertura lemmi non ambigui per classe

Classe	Lemmi	Token	% Lemmi	% Token
ANIMATE	4970	13866	32,7%	16,1%
ABSTRACT	2975	11154	52,3%	29,1%
LOCATION	1394	2939	61,5%	32,5%
OBJECT	1162	2317	69,1%	35,2%
EVENT	904	2160	75,1%	37,7%
O	759	1069	80,1%	39%

I restanti 3020 lemmi risultano ambigui, ma in misura diversa.

Una parte di questi (1936, il 12,7% del totale) sono in realtà ambigui solo se si considerano le istanze provenienti da documenti diversi.

Un altro aspetto da considerare per una valutazione sui livelli di ambiguità risultanti dall'annotazione è la preponderanza, in termini di frequenza, di un dato senso rispetto agli altri.

Considerato:
$s^\star = s \in S_l \mid \forall\, s' \in S_l\, occ\,(l, s) \geq occ\,(l, s')$ il senso più frequente di un lemma nel corpus, una misura di ambiguità per un lemma può essere definita come segue:

$$A\,(l) = \frac{1}{|S_l| - 1} \sum_{s \in S_l \setminus \{s^\star\}} \frac{occ\,(l, s)}{occ\,(l, s^\star)} \tag{2}$$

Così definita $A\,(l) \to 1$ se le occorrenze dei vari sensi del lemma risultano bilanciate, mentre $A\,(l) \to 0$ nel caso in cui uno dei sensi risulti nettamente prevalere sugli altri.

Raggruppiamo dunque i lemmi in esame in gruppi così definiti come nella formula 3, ottenendo quanto riportato in Tabella 8.

$$G(a, b) = |\{l \mid a < A(l) \leq b\}| \text{ con } a, b \in [0, 1] \tag{3}$$

I dati mostrano che nella maggior parte dei casi esiste un senso che ha il doppio o più delle occorrenze degli altri, e che quando ciò non è vero le occorrenze si distribuiscono in porzioni di testo disgiunte, che spesso implica sottodomini diversi.

Tabella 8:

		nello stesso documento		in documenti diversi		
a	**b**	**G(a,b)**	**%**	**G(a,b)**	**%**	**Δ**
0	0,1	163	0,215	341	0,153	0,06
0,1	0,2	145	0,191	357	0,160	0,03
0,2	0,3	111	0,146	253	0,113	0,03
0,3	0,4	74	0,097	283	0,127	0,03
0,4	0,5	107	0,141	389	0,174	0,03
0,5	0,6	29	0,038	65	0,029	0,01
0,6	0,7	36	0,047	87	0,039	0,01
0,7	0,8	29	0,038	66	0,030	0,01
0,8	0,9	19	0,025	28	0,013	0,01
0,9	1	46	0,061	364	0,163	**0,10**