

Introduzione al Trattamento Automatico del Linguaggio

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

Centro Interdisciplinare di Ricerche per la Computerizzazione
dei Segni dell'Espressione (CIRCSE)



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

DEFINIZIONI

Computational linguistics and natural language processing [...] are sometimes used interchangeably to describe the field concerned with the processing of human language by computers

- **Computational Linguistics** is used to describe research interested in answering linguistic questions using computational methodology
- **Natural Language Processing** describes research on automatic processing of human language for practical applications

Bender, Emily M. 2016. "Linguistic Typology in Natural Language Processing". Linguistic Typology 20(3), 645-660.

DEFINIZIONI

Testo e Computer, 2016

“L’obiettivo centrale della Linguistica Computazionale (LC) è quello di sviluppare modelli computazionali della lingua, cioè modelli del **funzionamento del linguaggio naturale** che possano essere tradotti in programmi eseguibili dal calcolatore e che consentano a quest’ultimo di acquisire le competenze necessarie per comunicare direttamente nella nostra lingua”



COSA VUOL DIRE STUDIARE IL FUNZIONAMENTO DEL LINGUAGGIO?

Il computer può essere usato per la gestione e l'analisi avanzata dei dati linguistici in formato digitale studiando, ad esempio:

- le costruzioni grammaticali
- la distribuzione della parole
- i cambiamenti semantici delle parole nel tempo
- le differenze linguistiche tra vari registri/autori/generi

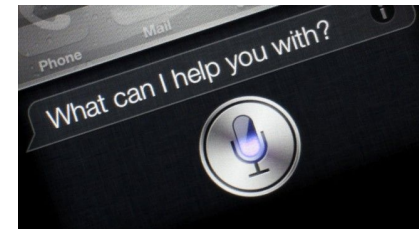
Fonetica	Studia la produzione e la percezione dei suoni
Fonologia	Studia il sistema mentale dei suoni
Morfologia	Studia la formazione e la struttura interna delle parole
Sintassi	Studia la struttura interna delle frasi
Semantica	Studia il significato delle parole o delle frasi
Pragmatica	Studia l'uso contestuale della lingua

MA...

Il computer, di per sé, **NON** conosce il linguaggio naturale!

Il **Trattamento Automatico del Linguaggio** (TAL) ha lo scopo di dotare il computer di conoscenze linguistiche, di creare macchine che capiscano (e addirittura riproducano) il linguaggio naturale, di sviluppare programmi che assistano l'essere umano in compiti (*task*) linguistici:

- riconoscimento automatico del parlato
- sintesi automatica della voce
- traduzione automatica
- analisi automatica del sentimento



PERCHÉ È UNA SFIDA

1. Ambiguità grammaticale

PAROLA	CATEGORIA GRAMMATICALE
C'	AVVERBIO/PRONOME
era	VERBO/NOME
una	ARTICOLO/PRONOME/NUMERALE
volta	NOME/VERBO (voltare)/VERBO (volgere)
un	ARTICOLO/NUMERALE
pezzo	NOME
di	PREPOSIZIONE
legno	NOME/VERBO

PERCHÉ È UNA SFIDA

2. Ambiguità sintattica: «una vecchia porta la sbarra»



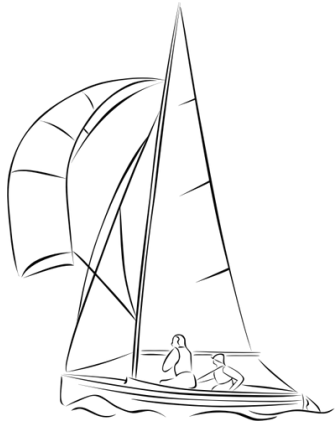
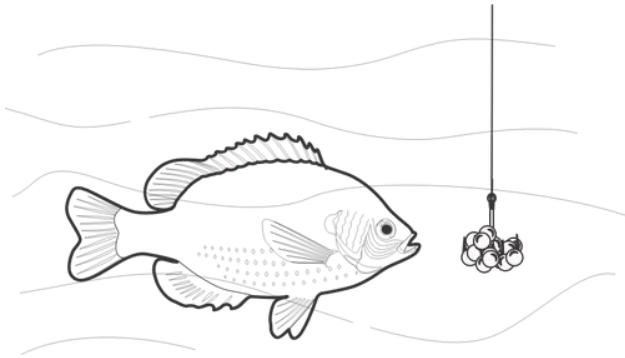
«una **vecchia** porta la sbarra»



«una vecchia **porta** la sbarra (la strada)»

PERCHÉ È UNA SFIDA

3. Ambiguità semantica: «*amo*» / «*navigare*»



PERCHÉ È UNA SFIDA

4. La lingua cambia

- Lingue classiche/storiche:

*Ahi quanto a dir qual era è cosa dura
esta selva selvaggia e aspra e forte
che nel pensier rinova la paura!*



- Lingue non-standard:

[#SanremoFunky](#) con [@elodie](#) e qualche considerazione sulla prima serata di [#Sanremo2020](#) 🎵 che sta per partire

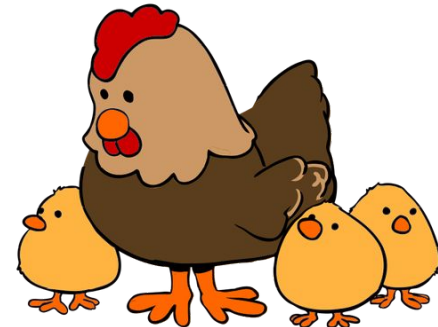
- Neologismi: *petaloso* / *Brexit*

PERCHÉ È UNA SFIDA

5. Espressioni multi-parola, ovvero «2 +2 non fa sempre 4»

Il loro significato non corrisponde alla combinazione lessicale delle parole che li compongono

- espressioni metaforiche: «*parlare dietro le spalle*»
- proverbi: «*si salvi chi può*»
- espressioni idiomatiche: «*conosco i miei polli*»



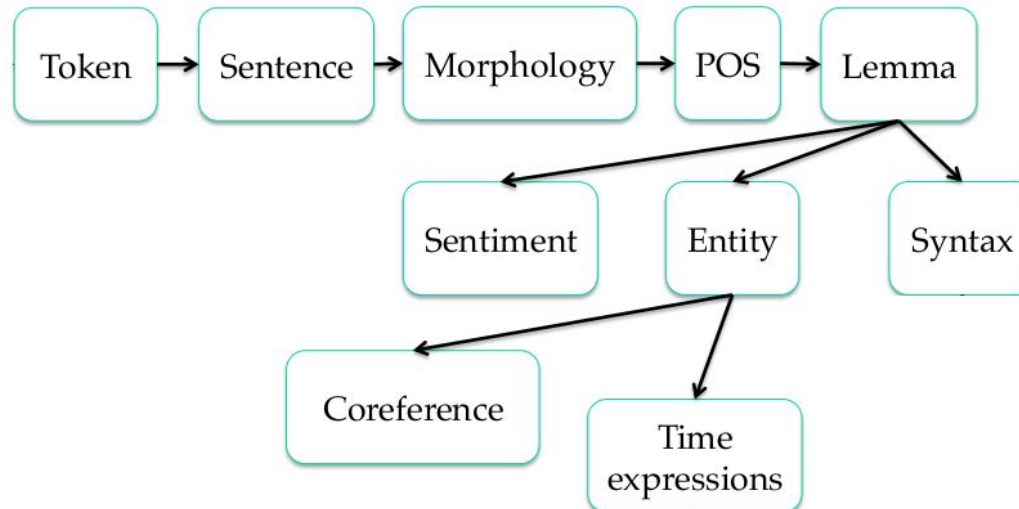
PERCHÉ È UNA SFIDA

6. Servono informazioni di contesto o di conoscenza del mondo
«Elsa e Anna sono sorelle»



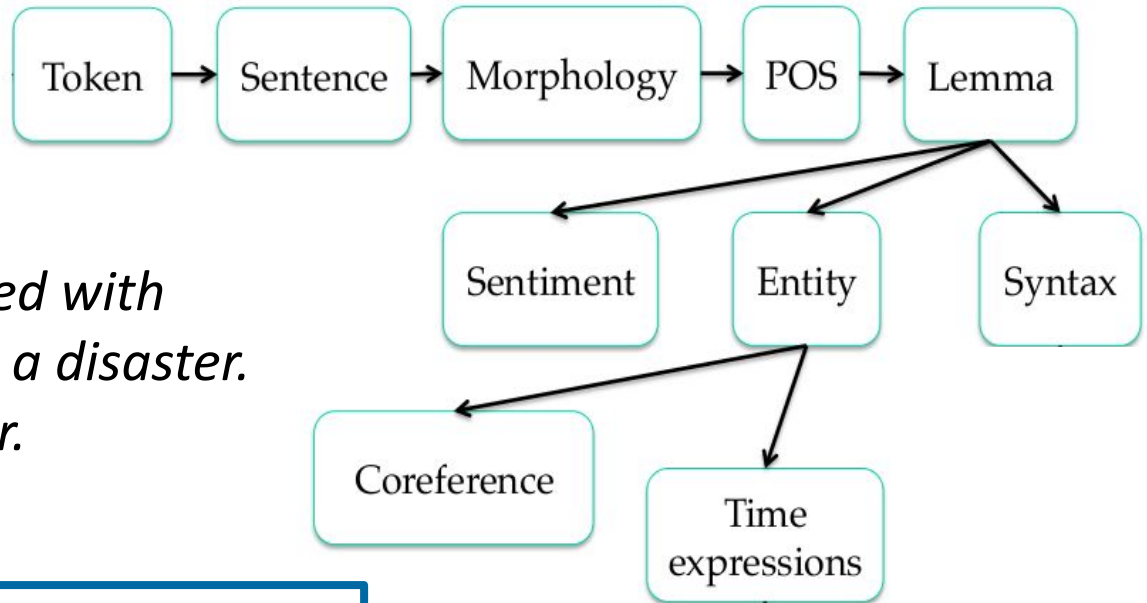
COME ANALIZZARE IL LINGUAGGIO

- Struttura a **PIPELINE**: catena i cui moduli descrivono ognuno un diverso livello di analisi linguistica e dove l'output di un modulo diventa l'input per il modulo successivo. Esempio:



Le analisi presentate nelle prossime slide sono
l'output della pipeline di Stanford CoreNLP

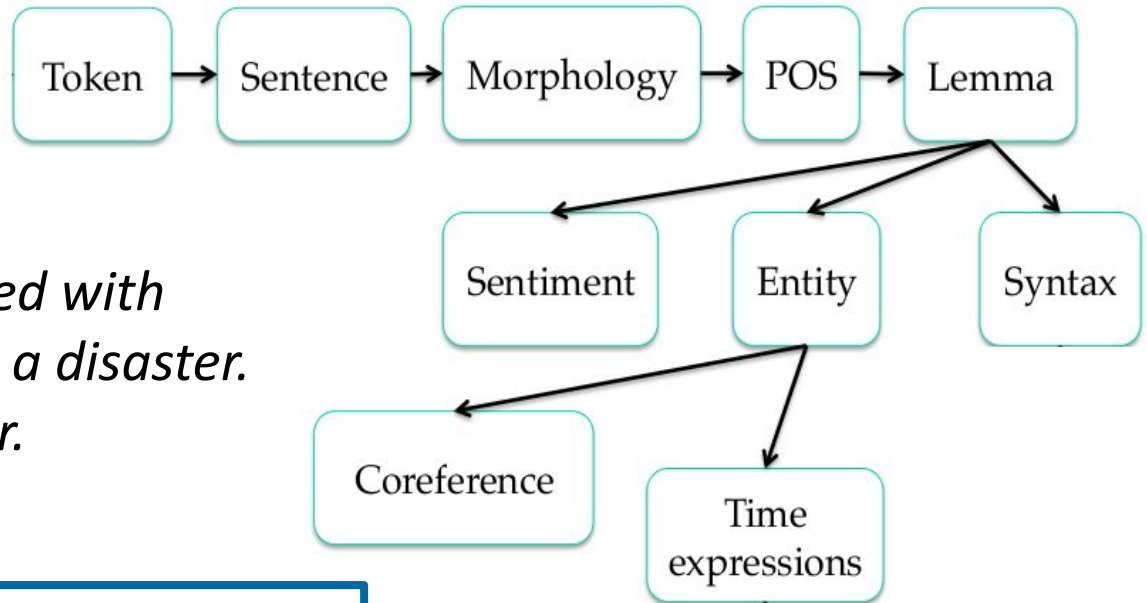
- demo online: <http://corenlp.run/>



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
 Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH

	WRB	PRP	VBP	WP	VBD	IN	JJ	NNP	NN	,	PRP	VBD	DT	NN	.
1	When	you	see	what	happened	with	crooked	Hillary	today	,	it	was	a	disaster	.
2	DT	NN	.												
	A	disaster	.												
3	PRP	VBD	DT	NN	.										
	She	had	a	disaster	.										



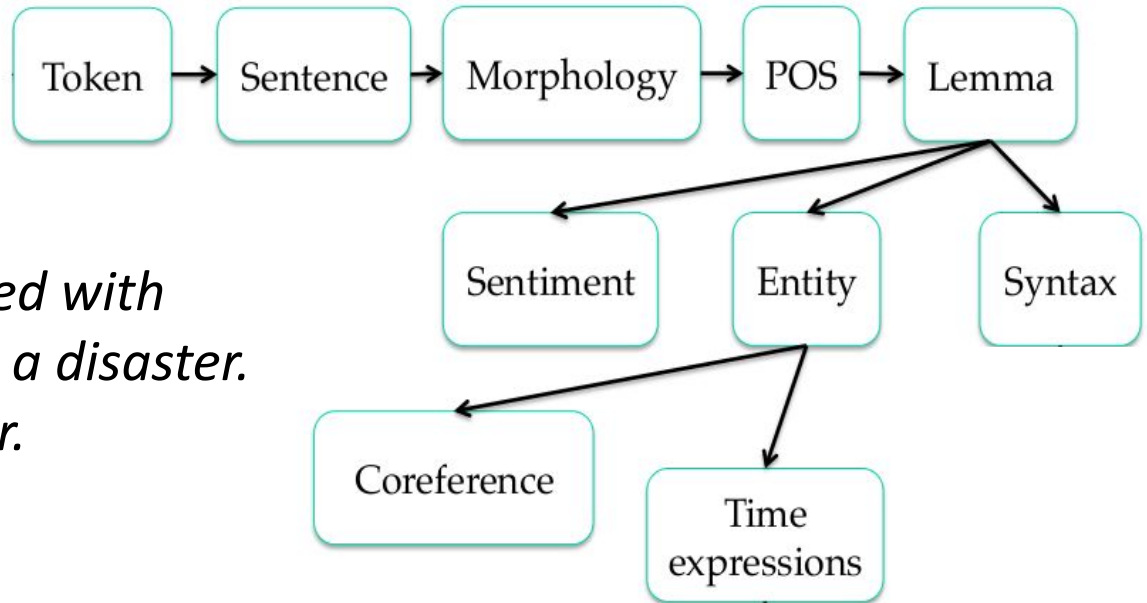
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH

C'era una volta un pezzo di legno.

C'era | una | volta | un | pezzo | di | legno.

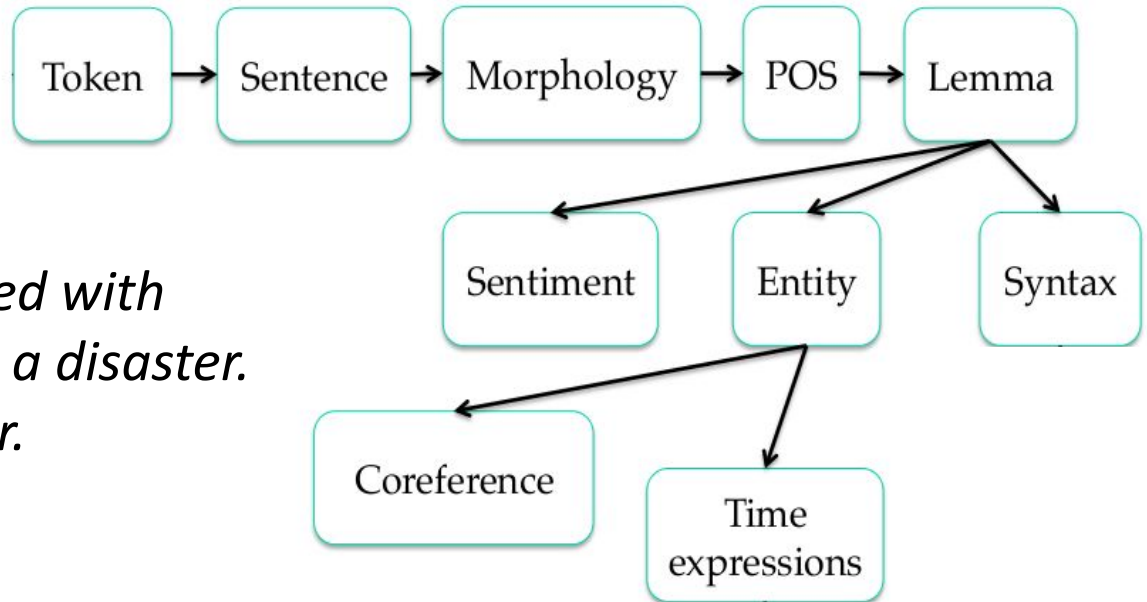
C' | era | una | volta | un | pezzo | di | legno | .



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
 Trump, 2016-08-05

MORPHOLOGY

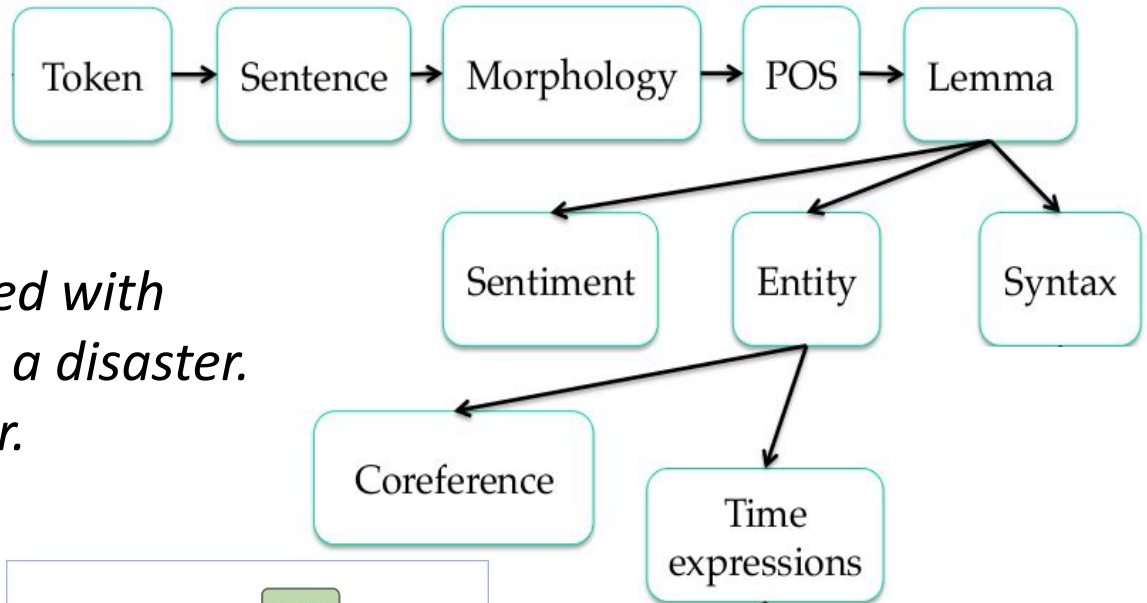
when+conj	you+pron	see+v+indic+pres+no3sing	what+adj+zero	happen+v+indic+past	with+prep	crooked+adj+zero	NULL	today+adv	NULL
When	you	see	what	happened	with	crooked	Hillary	today	,
it+pron	be+v+indic+past	a+art	disaster+n+sing	disaster	disaster
it	was	a	disaster
a+art	disaster+n+sing
A	disaster
she+pron	have+v+indic+past	a+art	disaster+n+sing	disaster	disaster
She	had	a	disaster



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

LEMMA

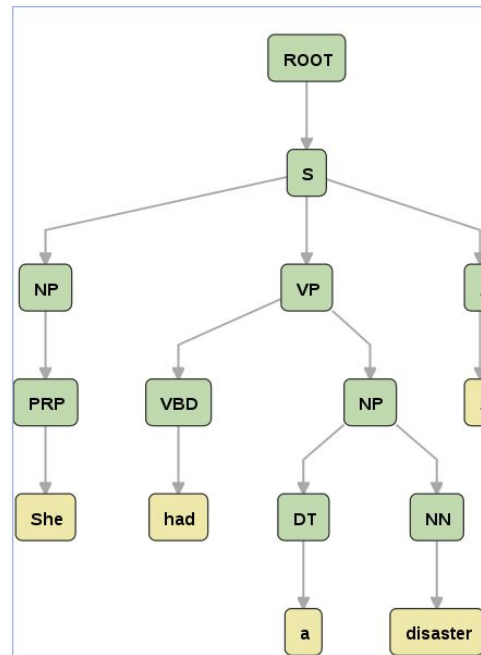
- 1 when you see what happen with crooked Hillary today , it be a disaster .
- 2 a disaster .
- 3 she have a disaster .

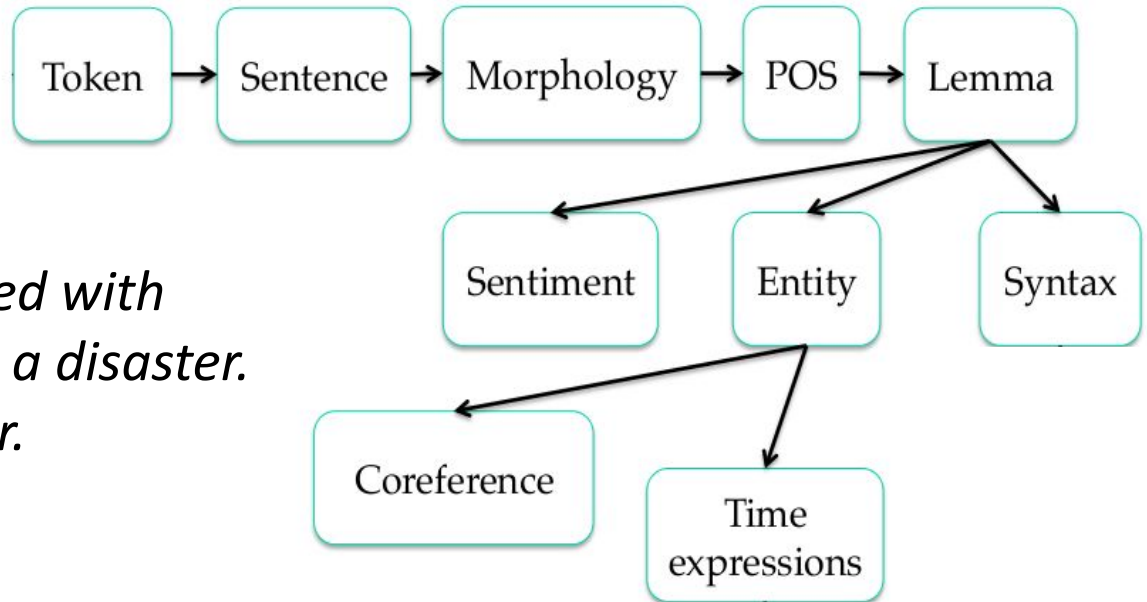


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SYNTAX / PARSING

- a costituenti

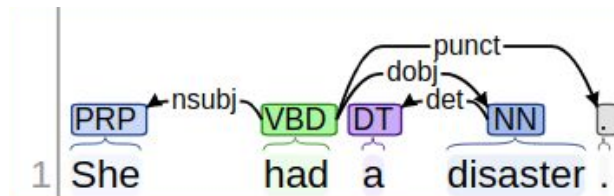


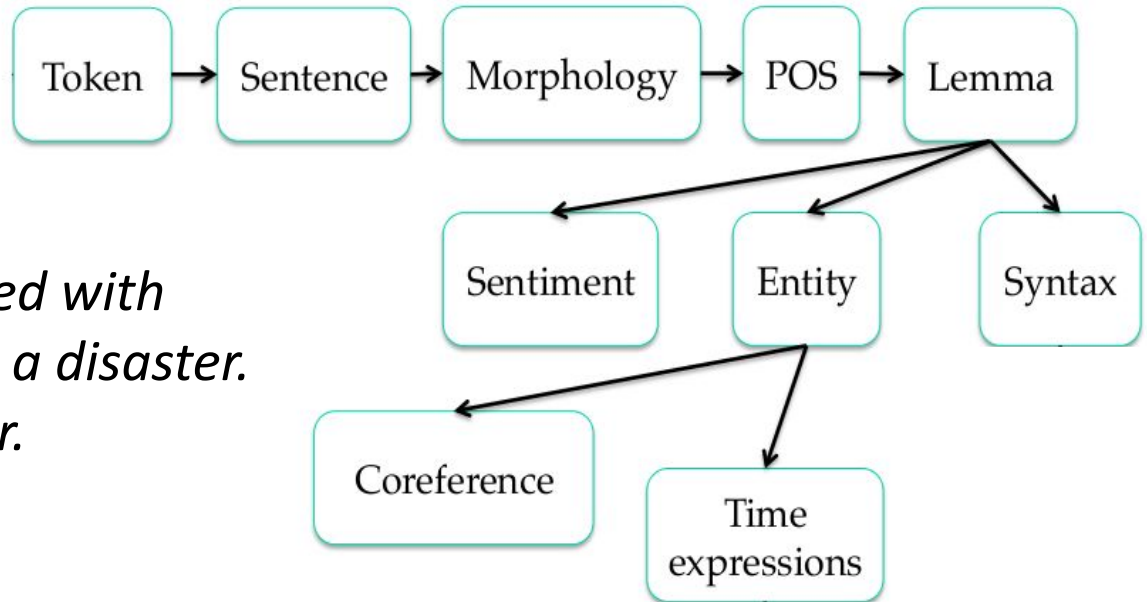


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SYNTAX / PARSING

- a dipendenze

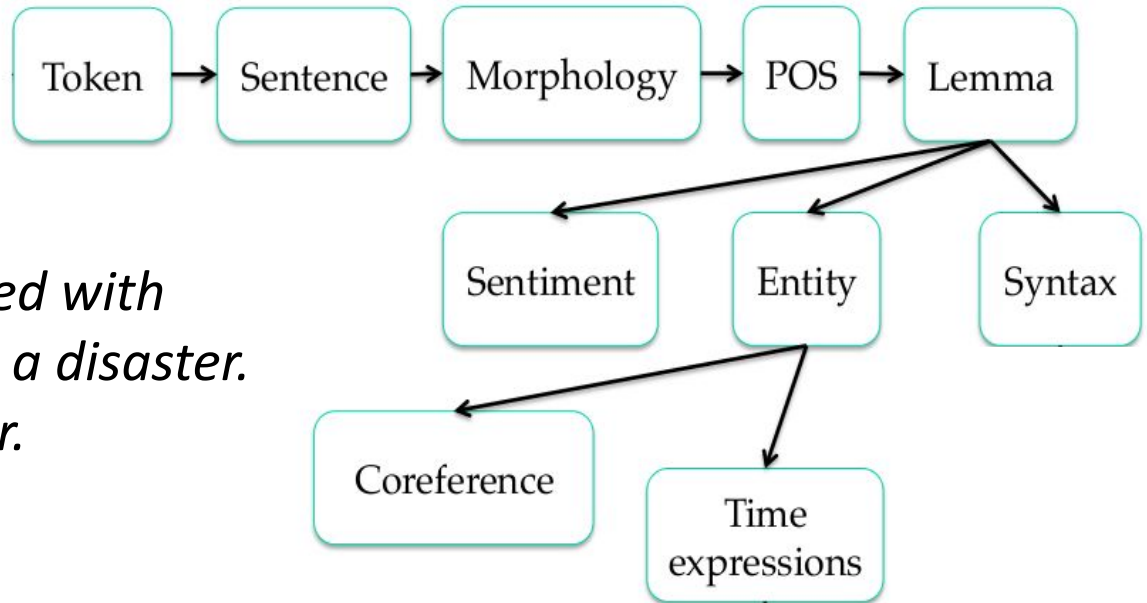




When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

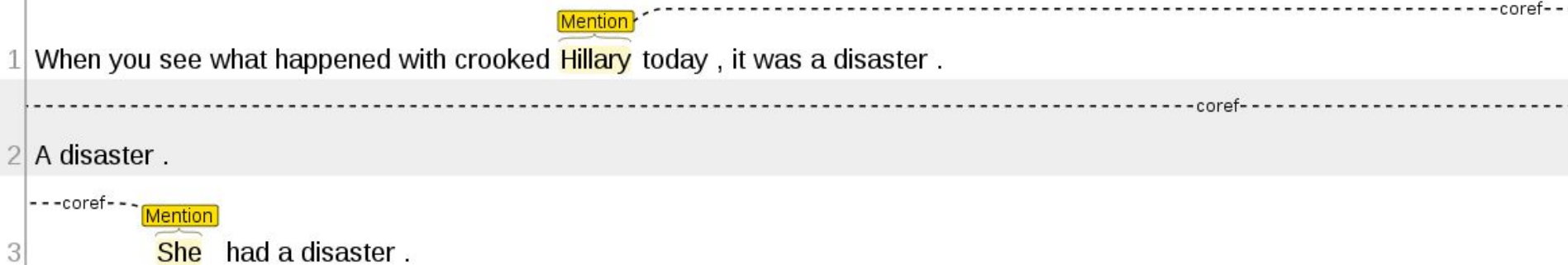
ENTITY

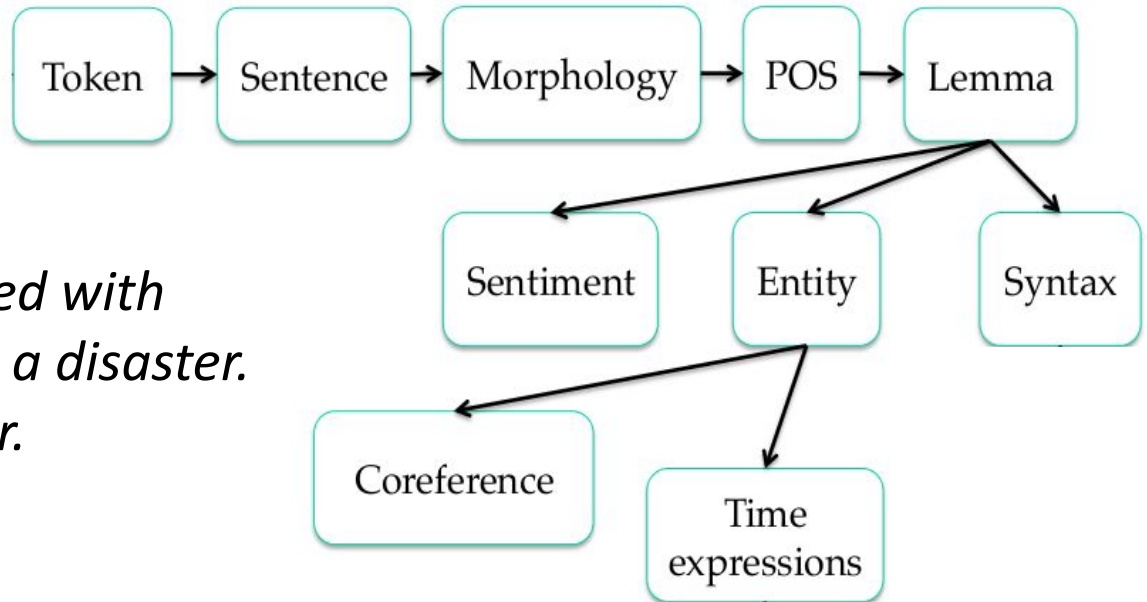
- 1 When you see what happened with crooked PER Hillary today , it was a disaster .
- 2 A disaster .
- 3 She had a disaster .



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

COREFERENCE



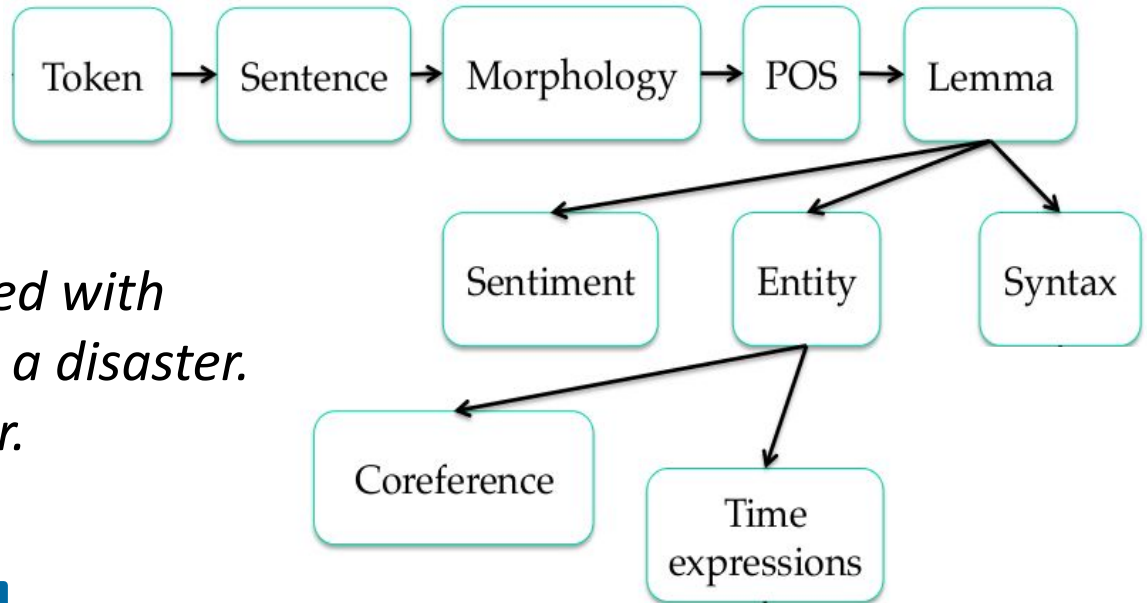


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TIME EXPRESSIONS

2016-08-05

- 1 When you see what happened with crooked Hillary today , it was a disaster .
- 2 A disaster .
- 3 She had a disaster .



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SENTIMENT

		NEGATIVE
1	When you see what happened with crooked Hillary today , it was a disaster .	
2	A disaster .	VERY NEGATIVE
3	She had a disaster .	NEGATIVE

COME SI SVILUPPA UN MODULO TAL

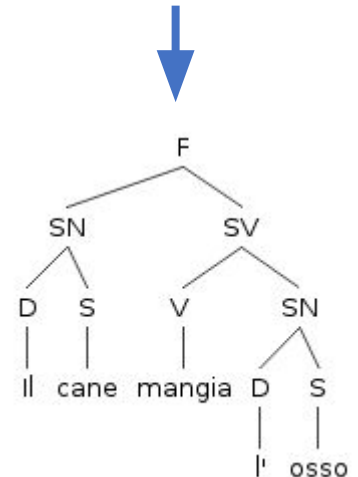
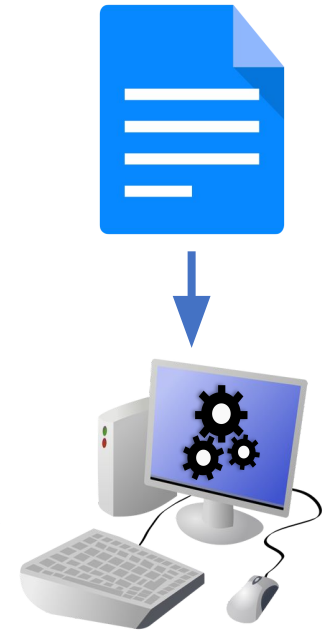
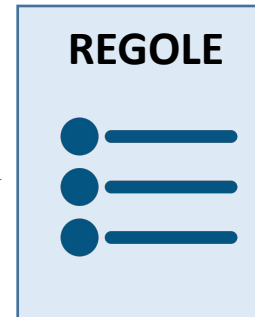
LOOKUP LIST

- Sistema che riconosce solo le parole memorizzate nei suoi elenchi detti “gazetteers”
- Vantaggi: semplice, veloce, facile da utilizzare
- Svantaggi: la raccolta e il mantenimento degli elenchi richiede tempo, gli elenchi non gestiscono tutte le possibili varianti delle parole e non possono risolvere l'ambiguità, nessun tipo di inferenza

LISTA_VALUTE	LISTA_CITTÀ
Euro, dollaro, dollari, sterlina, sterline, \$, €...	http://download.geonames.org/export/dump/

COME SI SVILUPPA UN MODULO TAL

SISTEMI A REGOLE (RULE BASED)



PRO

- basato su evidenze linguistiche
- preciso

CONTRO

- difficile da estendere o da adattare a nuovi domini
- richiede tempo per essere sviluppato

COME SI SVILUPPA UN MODULO TAL

SISTEMI A REGOLE (RULE BASED)

- Esempio: Part-of-Speech tagging:

1) assegnazione ad ogni parola di tutti i possibili PoS usando un dizionario

VERB		NOUN
«paghiamo	ART	VERB
	il	conto»

2) applicazione delle regole per rimuovere etichette ambigue

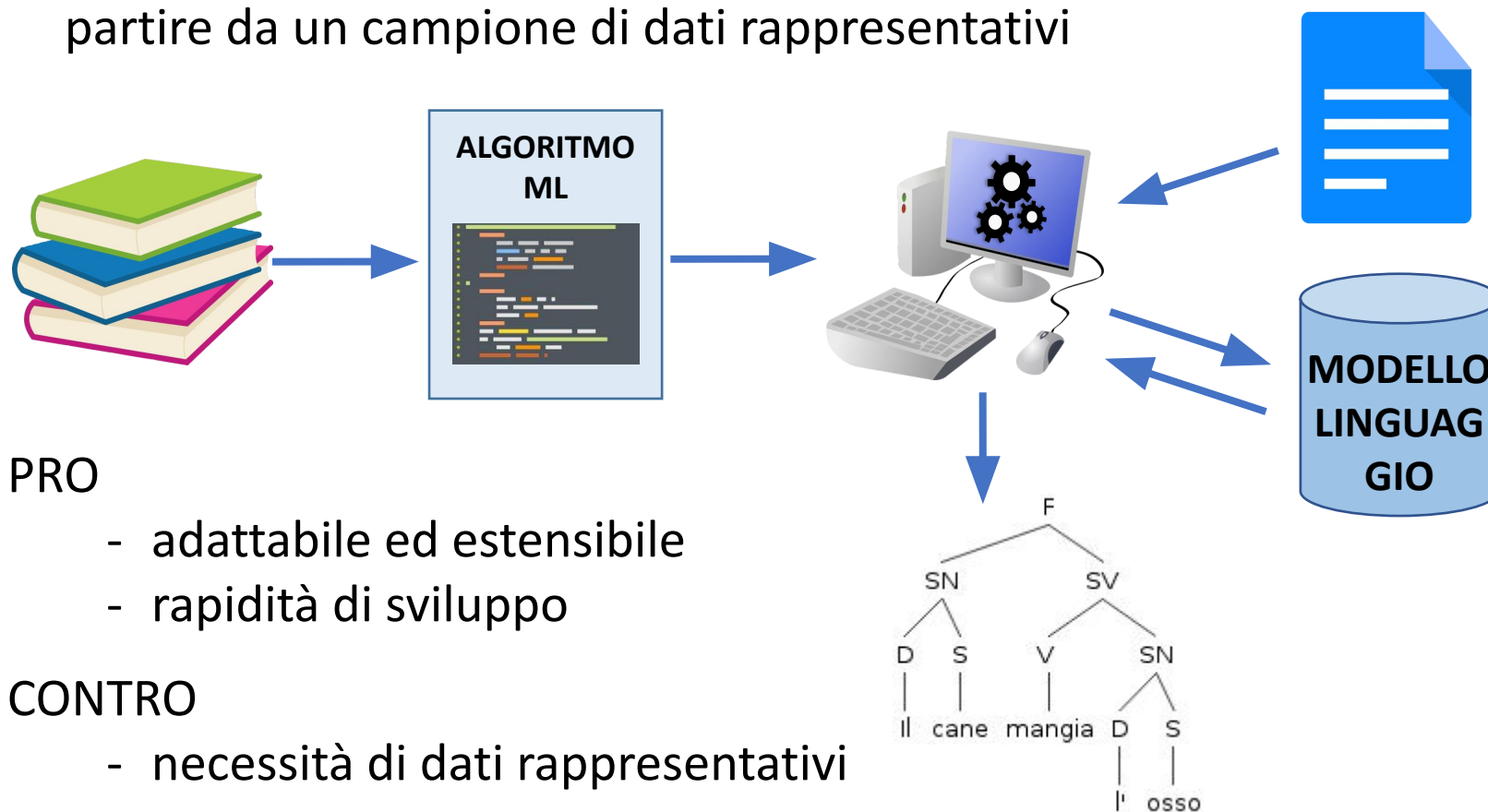
- «rimuovere VERB se in alternativa con NOUN e preceduto da ART»

VERB		NOUN
«paghiamo	ART	VERB
	il	conto»

COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- algoritmi che permettono al computer di imparare a svolgere un task a partire da un campione di dati rappresentativi



COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- **3 tipi principali di algoritmi di ML**

1. **NON SUPERVISIONATI:** non necessitano di un corpus annotato a mano per creare il modello
2. **SUPERVISIONATI:** utilizzano un corpus annotato a mano per la creazione dei modelli
3. **SEMI-SUPERVISIONATI:** combinano informazioni derivanti sia da corpora annotati che da dati non annotati

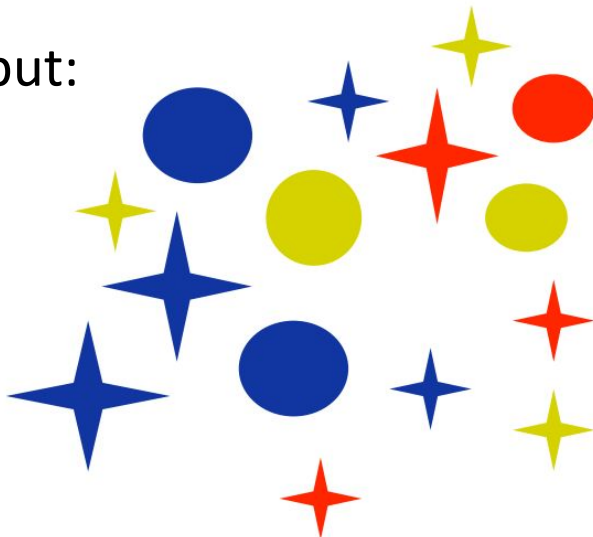
COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

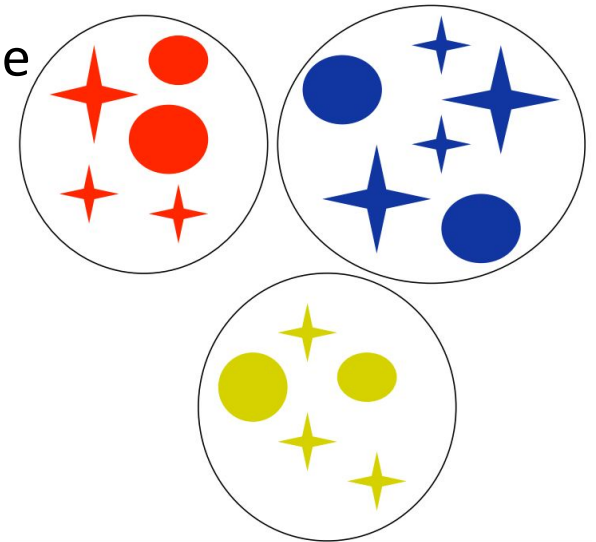
- **ML NON SUPERVISIONATO, esempio**

- CLUSTERING: raggruppamento dell'input in base a una qualche relazione di similitudine tra i dati

Input:



Output in base
al colore:



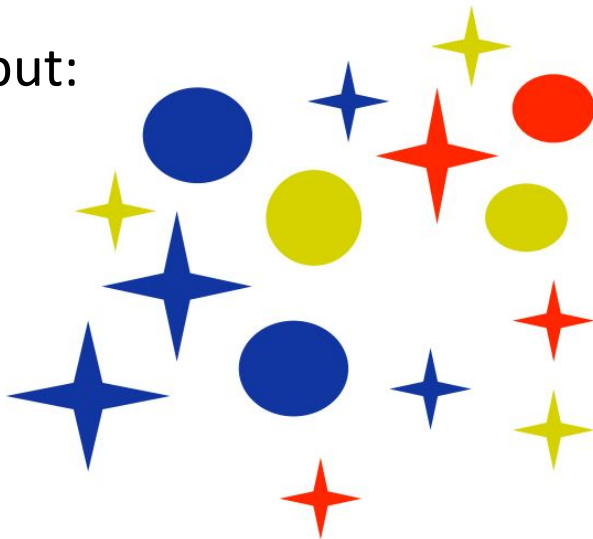
COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

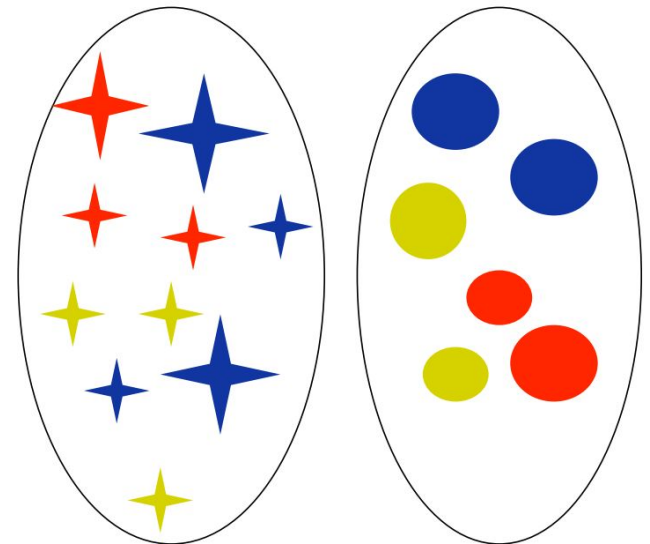
- **ML NON SUPERVISIONATO, esempio**

- CLUSTERING: raggruppamento dell'input in base a una qualche relazione di similitudine tra i dati

Input:



Output in
base alla
forma:



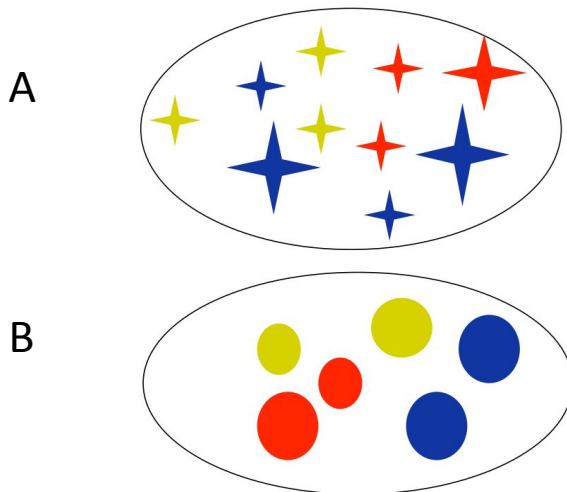
COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

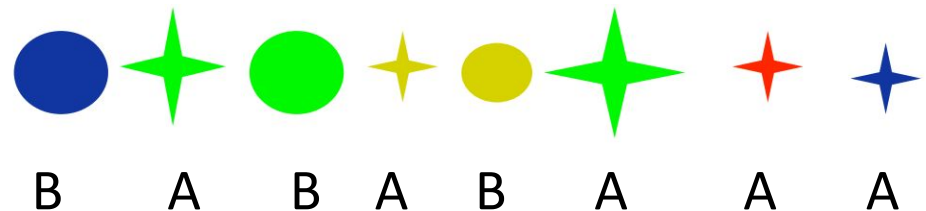
- **ML SUPERVISIONATO, esempio**

- **CLASSIFICAZIONE:** dato un insieme di classi predefinite determinare a quale classe appartiene una certa entità

Input (training):



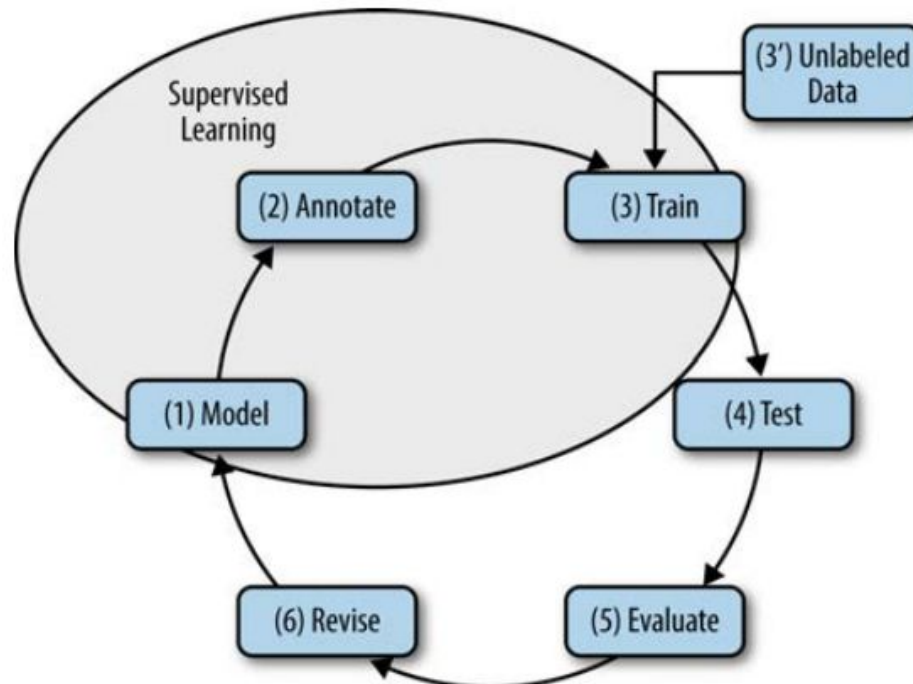
Classificazione di nuovi dati (test):



COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- ML SUPERVISIONATO



Il ciclo MATTER

(Pustejovsky and Stubbs (2012) "Natural Language Annotation for Machine Learning". O'Reilly Media.)

COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- **ML SUPERVISIONATO**

- Il ciclo MATTER:

- **Model**: descrizione teorica di un fenomeno linguistico
- **Annotate**: annotazione del corpus con uno schema di annotazione basato sul modello
- **Train**: addestramento di un algoritmo di ML sul corpus annotato
- **Test**: test del sistema addestrato su un nuovo campione di dati
- **Evaluate**: valutazione delle performance del sistema
- **Revise**: revisione del modello e dello schema di annotazione

COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- **ML SUPERVISIONATO**
- **ANNOTAZIONE**
 - aggiunta di informazioni (linguistiche) al testo tramite etichette (*tag*)
 - copre ogni aspetto dell'analisi linguistica
 - rende esplicita e analizzabile dal computer la struttura linguistica implicita nel testo
- **SCHEMA DI ANNOTAZIONE**
 - repertorio di categorie per l'annotazione: lista di tag e attributi
- **LINEE GUIDA DI ANNOTAZIONE**
 - documento in cui viene spiegato il *modo* in cui l'annotazione è proiettata sul testo

COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- **ML SUPERVISIONATO**

- **Dati necessari:**

- di training (*training set*): dati annotati per l'addestramento del modello
- di test (*test set*): dati NON annotati, diversi da quelli di training, su cui applicare il modello addestrato
- di valutazione (*gold standard*): dati del test annotati su cui valutare le performance del modello addestrato

COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- ML SUPERVISIONATO

- Esempio: ***Sentiment Polarity Classification***

subj	Subjectivity: possible values are 0 and 1. A subjective tweet will have subj = 1; an objective tweet subj = 0.
opos	Positive <i>overall</i> polarity: possible values are 0 and 1. A tweet exhibiting positive polarity will have opos = 1; a tweet without positive polarity will have opos = 0.
oneg	Negative <i>overall</i> polarity: possible values are 0 and 1. A tweet exhibiting negative polarity will have neg = 1; a tweet without negative polarity will have neg = 0.
iro	Irony: possible values are 0 and 1. A tweet with an ironic twist will have iro = 1, otherwise iro = 0.
lpos	Positive <i>literal</i> polarity: possible values are 0 and 1. A tweet exhibiting positive <i>literal</i> polarity will have pos = 1; tweet without positive <i>literal</i> polarity will have pos = 0.
lneg	Negative <i>literal</i> polarity: possible values are 0 and 1. A tweet exhibiting negative <i>literal</i> polarity will have neg = 1; tweet without negative <i>literal</i> polarity will have neg = 0.

COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- ML SUPERVISIONATO

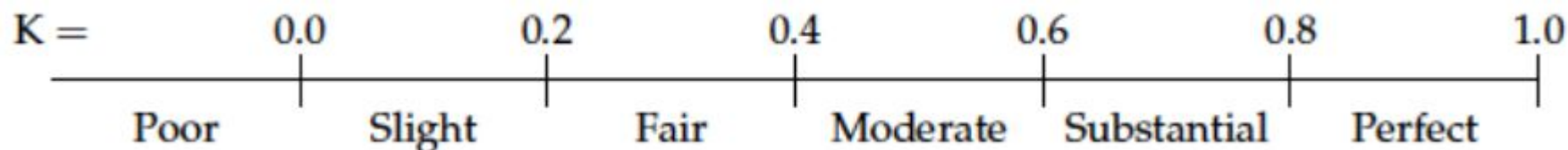
- Esempio: *Sentiment Polarity Classification*

subj	opos	oneg	iro	lpos	lneg	description and explanatory tweet in Italian
0	0	0	0	0	0	objective <i>l'articolo di Roberto Ciccarelli dal manifesto di oggi</i> http://fb.me/1BQVy5Wak
1	0	0	0	0	0	subjective with neutral polarity and no irony <i>Primo passaggio alla #strabrollo ma secondo me non era un iscritto</i>
1	1	0	0	1	0	subjective with positive polarity and no irony <i>splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura</i> http://t.co/GWoZqbxAuS
1	0	1	0	0	1	subjective with negative polarity and no irony <i>Monti, ripensaci: l'inutile Torino-Lione inguaia l'Italia: Tav, appello a Mario Monti da Mercalli, Cicconi, Pont...</i> http://t.co/3CazKS7Y
1	1	1	0	1	1	subjective with both positive and negative polarity (mixed polarity) and no irony <i>Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme"</i> http://t.co/kIKnbFY7
1	1	0	1	1	0	subjective with positive polarity, and an ironic twist <i>Questo governo Monti dei paschi di Siena sta cominciando a carburare; speriamo bene...</i>

COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

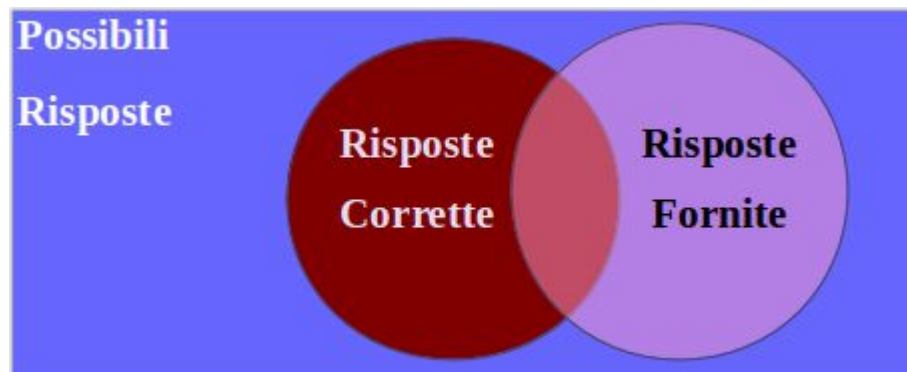
- **ML SUPERVISIONATO**
- **Inter-Annotator Agreement (IAA)** = accordo tra almeno 2 annotatori sullo stesso testo
 - consistenza dell'annotazione
 - plausibilità cognitiva del modello
 - un ampio accordo tra gli annotatori è considerato garanzia della validità di tale schema e dei dati annotati
 - K di Cohen (annotatori = 2) o di Fleiss (annotatori > 2)



COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- **ML SUPERVISIONATO**
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - confronto dell'output del modello sui dati di test con il gold standard



COME SI SVILUPPA UN MODULO TAL

Sistemi di apprendimento automatico – MACHINE LEARNING (ML)

- **ML SUPERVISIONATO**
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: ACCURACY

$$\text{ACCURACY} = \frac{\text{\#risposte corrette}}{\text{\#risposte fornite}}$$

Esempio:

- 150 frasi annotate nel test
- 120 frasi annotate con sentiment corretto
- accuracy = $120/150 = 0,8$ (80%)



GRAZIE!

Email: rachele.sprugnoli@unicatt.it

Twitter: [@RSprugnoli](https://twitter.com/RSprugnoli)



COME SI SVILUPPA UN MODULO TAL


- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** matrice di confusione

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: **PRECISION**, misura il rapporto tra le entità correttamente riconosciute dal sistema ed il totale delle entità riconosciute

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

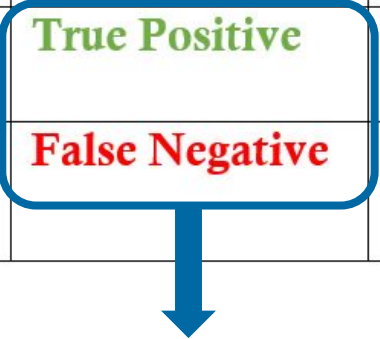


$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: **RECALL**, misura il rapporto tra le entità correttamente riconosciute dal sistema ed il totale delle entità corrette

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: **F-MEASURE**, media armonica tra precision e recall

$$\text{F-MEASURE} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - Esempio:

		ACTUAL (gold standard)	
		Positive	Negative
PREDICTED (test set)	Positive	70 (TP)	15 (FP)
	Negative	30 (FN)	45 (TN)

- Precision: $70 / (70+15) = 70 / 85 = 0,82$

- Recall: $70 / (70+30) = 70 / 100 = 0,70$

- F-measure: $2 * 0,82 * 0,7 / (0,82 + 0,70) = 0,75$