The Italian NLP Filter

Paolo Allegrini, Simone Marchi,
Simonetta Montemagni

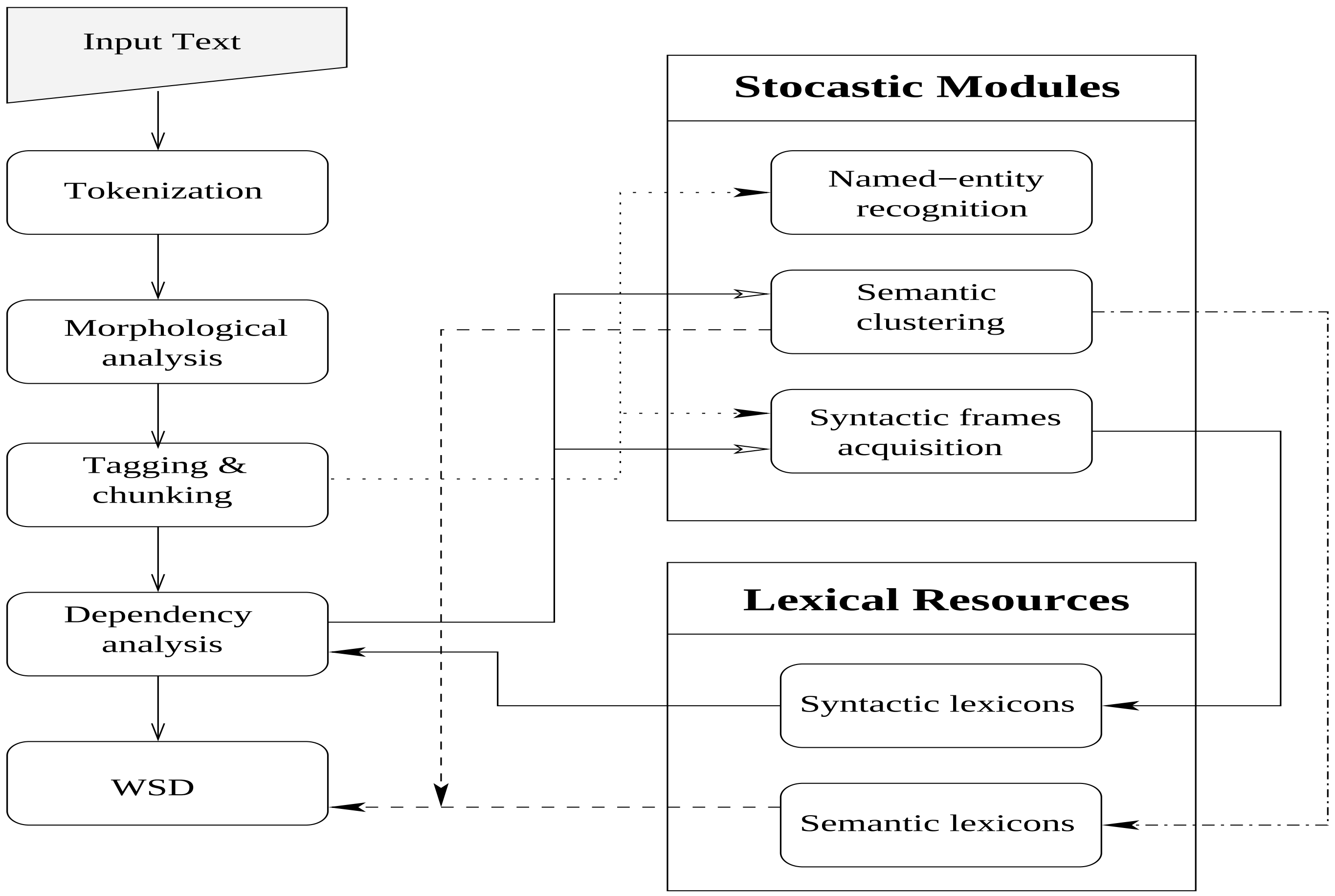
Istituto di Linguistica Computazionale del Consiglio
Nazionale delle Ricerche, Area della Ricerca di Pisa,
via Alfieri 1, S. Cataldo, 56124 Pisa, Italy

allegrip@ilc.cnr.it simone.marchi@ilc.cnr.it
simo@ilc.cnr.it

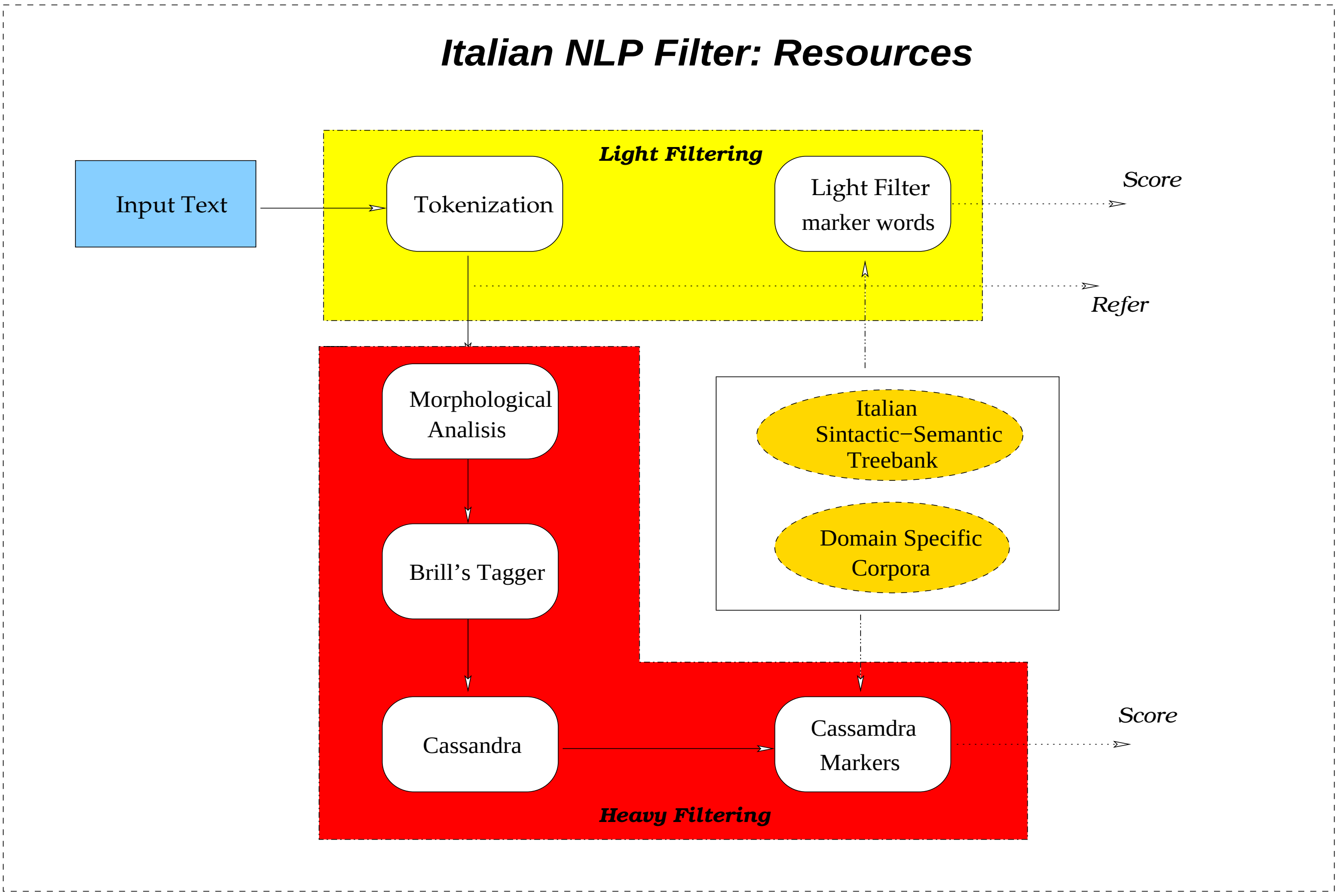
Summary

- Resources available (online and offline)
- Architecture
- Light filter: Off-line NLP processing and regular espressions
- Heavy filter: Online NLP processing and CASSANDRA
- Evaluation (in house and at Telefonica)
- Conclusions

ILC NLP resources for content filtering



ILC NLP filter architecture



The Light Filter

- Saliency score acquired from a morpho-syntactically annotated and lemmatised training corpus of 300.000 word tokens

- List of "marker"-words (30 lemmata) $\rightarrow$ *indirect lemmatisation strategy*

- Tokenised text, segmented into text windows of 100 words

- Score = maximum local frequency relevant words in text windows

Recognizing salient words needs pos-tagging and lemmatisation.

Strategy: On-line Augmented Tokenization and off-line morphological generation.

The Heavy Filter

Operates on morpho-syntactically annotated and lemmatised text

Background resources:

longer list of "marker"-words (*including ambiguous ones])*
a corpus of typical word co-occurrence patterns
clusters of domain-specific semantically similar words

Background info automatically extracted from a training corpus of 1.660.000 word tokens through use of advanced NLP techniques (dependency analysis)

Implementation of a new entropy-based classification technique,called CASSANDRA (Complex Analysis of Sequences via Scaling AND Randomness Assessment)

Good results also on a corpus of erotic stories

Why an NLP-based Approach? -I-

Salience $s \equiv \frac{f_{ES} - f_{TB}}{f_{ES} + f_{TB}}$

we have for forms the following list:
(increasing s for words $\in TB \bigcap ES$)

```
                  ...
vedevo 0.9375 ginocchia 0.939394 saliva 0.939394
secchio 0.939394 tu 0.939394 spinse 0.941176
bocca 0.941691 lascio' 0.944444 reggiseno 0.944444
rispose 0.947368 stai 0.948718 uccello 0.95
fianchi 0.95122 riusciva 0.95122 jeans 0.953488
mie 0.956522 inizio' 0.960784 clitoride 0.961538
slip 0.961538 cosce 0.962617 vagina 0.962963
farmi 0.964286 stavo 0.964286 fino 0.965957
capezzoli 0.966667 Roberta 0.969349 Laura 0.971831
schiena  0.975 cosi 0.977273 seni 0.978022
Elena 0.984674 labbra 0.98895
```

Why an NLP-based Approach? -II-

... MORBIDO#A 0.857143 SUDATO#A 0.857143 COPERTO#A 0.866667 DUBBIO#A 0.866667 SORPRESO#A 0.894737 SODO#A 0.916667 UMIDO#A 0.923077 LEGATO#A 0.948718 ECCITATO#A 0.978947 TUTTO#A 0.994723

... DENTRO#B 0.851852 A_FATICA#B 0.857143 ATTENTAMENTE#B 0.875 LENTAMENTE#B 0.877301 VELOCEMENTE#B 0.911111

... PUTTANA#S 0.875 REGGISENO#S 0.891892 PORCO#S 0.904762 VAGINA#S 0.927273 CUSCINO#S 0.928571 PENETRAZIONE#S 0.935484 BOCCA#S 0.937143 CAPEZZOLO#S 0.952941 CALZA#S 0.954545 JEANS#S 0.954545 CLITORIDE#S 0.961538 SLIP#S 0.961538 SCHIENA#S 0.975309 LABBRO#S 0.989247

SOPRA#E 0.842105 DENTRO#E 0.901786

... ALZARE#V 0.884393 ASSAPORARE#V 0.888889 INFILARE#V 0.893805 SPORGERE#V 0.894737 VERGOGNARE#V 0.894737 ABBASSARE#V 0.897436 COLARE#V 0.913043 DIVARICARE#V 0.928571 DISTENDERE#V 0.945946 INGOIARE#V 0.955556 PENETRARE#V 0.965517 GEMERE#V 0.966667 ACCAREZZARE#V 0.972414 BACIARE#V 0.972789 ECCITARE#V 0.987654

salient words in $ES \bigcap TB$ *can still be used for filtering*

crude dimension-reduction: projection into one dimension

Mathematical Foundation of Heavy filter -I-

$P(x,t)$ is the prob. of finding x events in a segment of t words, chosen randomly. In other words, we put a 1 if a "marker" word is met, a 0 otherwise. $\rightarrow$ sequence $\{\xi_i\}$ as, e.g.
1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 1 1 0

Then x is defined as $$x_n(l) = \sum_{i=l}^{l+n} \xi_i$$

Considering $n \rightarrow t$, $x_n - \langle x_n \rangle \rightarrow x$, ergodicity *implies* scaling, *namely*

$$P(x,t) = \frac{1}{t^\delta} F\left(\frac{x}{t^\delta}\right) \tag{1}$$

$$S(t) = -\int_{-\infty}^{\infty} P(x,t) \ln\left[P(x,t)\right] \mathrm{d}x \tag{2}$$

$$S(t) = A + \delta \ln(t) \tag{3}$$

$$\langle x^2 \rangle \propto t^{2\delta}, \ \langle (\xi - \bar{\xi})(\xi(t) - \bar{\xi}) \rangle \propto t^{2\delta-2} \tag{4}$$

Mathematical Foundation of Heavy filter -II- *truncated Lévy diffusion process*

Events time distance t distribuited as

$$\psi(t) = (\mu - 1)\frac{T^{\mu-1}}{(t+T)^{\mu}}. \tag{5}$$

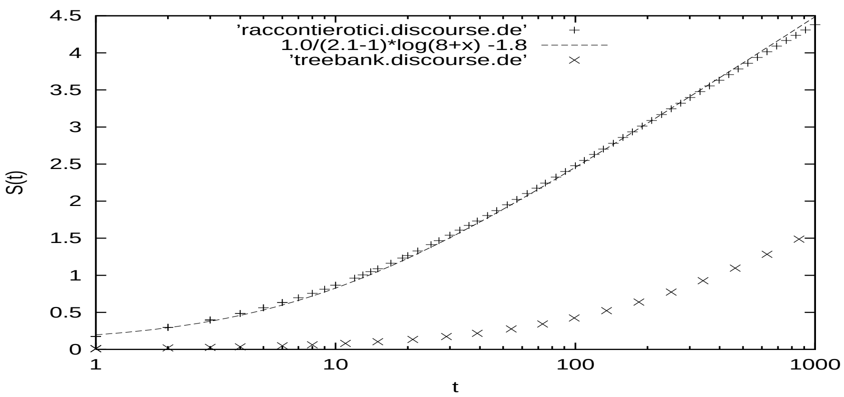
The theory based on CTRW and GCLT yields a (truncated) Levy PDF. DE detects the approximate scaling of the central part.

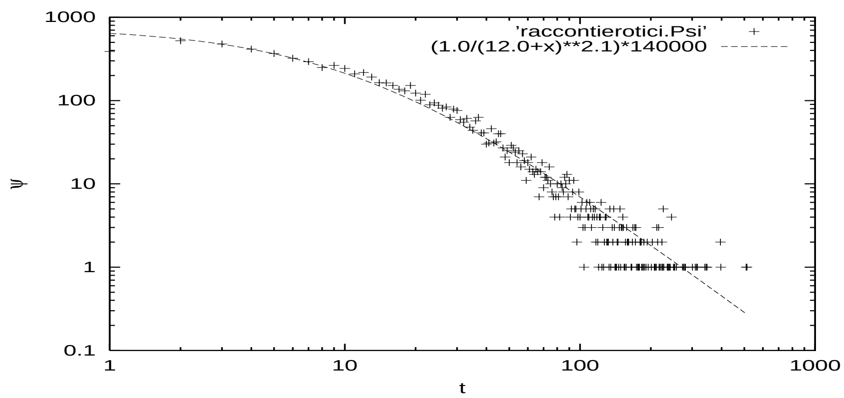
$$\delta = \frac{1}{\mu - 1} \text{ if } 2 < \mu < 3,\ \delta = 0.5 \text{ if } \mu > 3 \tag{6}$$

The theory rests on the dichotomous assumption.
It rests on uncorrelated waiting times between events.
Under these conditions each event carries a unit of information.

If in a corpus we find a marker (e.g. a list of words) such that $\delta \approx 1/(\mu - 1)$ then this marker is informative in that corpus.

Testing the Hypotesis





salient words seem to fit the information test.

Complex Analysis of Sequences via Scaling AND Randomness Assessment (CASSANDRA)

- We define a local entropy as a function of the position along the sequence of a "large" window

- We use a local complexity indicator:

$$\delta(T) = \frac{1}{N}\sum_{t=2}^{N}[S(t) - 1/2\ln(t) - S(1)] \tag{7}$$

if $\delta(x) = 0 \rightarrow$ no correlation

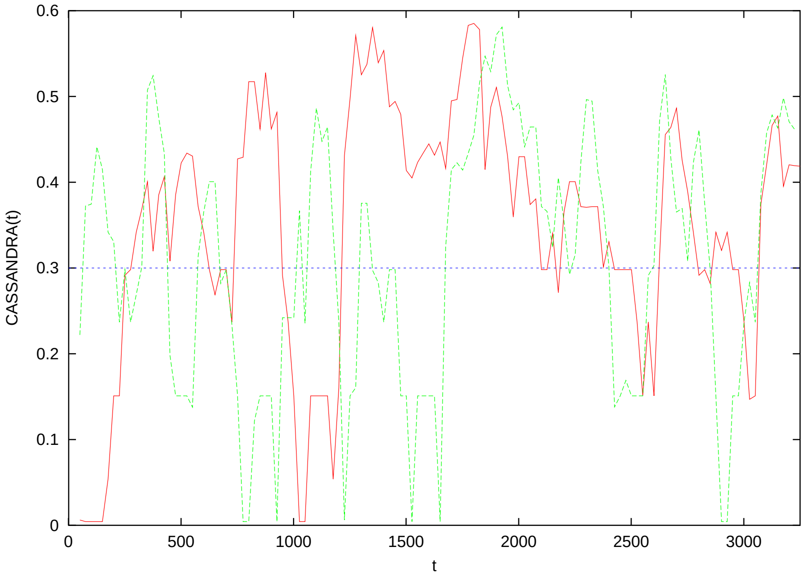
if $\delta(x) > 0 \rightarrow$ correlation (persistence)

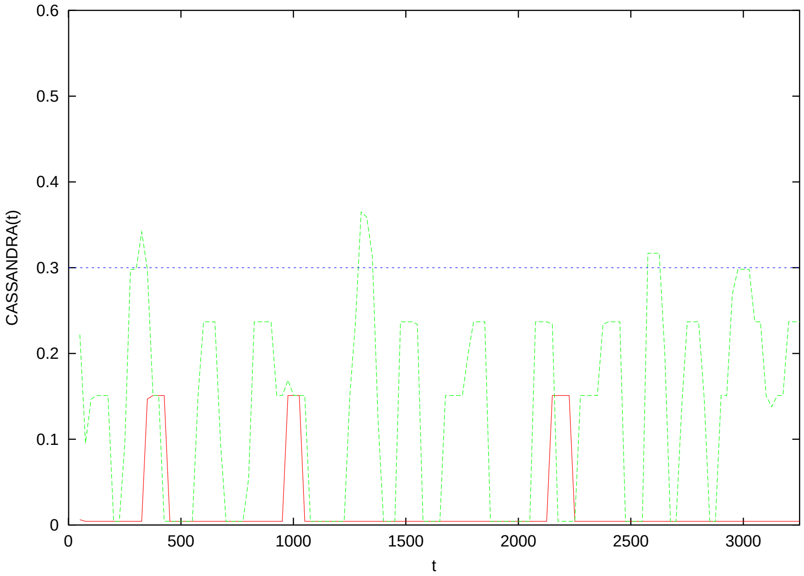
if $\delta(x) < 0 \rightarrow$ correlation (anti-persistence)

This method combines local frequency properties and correlational ones. In a sense it is similar to the wavelet transform.

we perform the analysis on training, control and test corpora, with a window-size of 100 words.

CASSANDRA at work





0.6
0.5
0.4
0.3
0.2
0.1
0
CASSANDRA(t)
0
500
1000
1500
2000
2500
3000
t

Italian NLP Filter Results

ILC evaluation for light filter

Predicted Actual	Harmful	Harmless	Unknown	Total
Harmful	3143	131	228	3502
Harmless	6	4010	179	4195
Total	3149	4141	407	7697
Precision	0.998	0.968		
Recall	0.897	0.956		
F-Measure	0.948	0.962		

ILC evaluation for light and heavy filter

Predicted Actual	Harmful	Harmless	Unknown	Total
Harmful	3181	165	156	3502
Harmless	15	4111	69	4195
Total	3196	4276	225	7697
Precision	0.995	0.961		
Recall	0.908	0.980		
F-Measure	0.952	0.970		

Telefónica evaluation (Image + NLP)

Predicted Actual	Harmful	Harmless	Unknown	Total
Harmful	7053	258	189	7500
Harmless	156	7126	218	7500
Total	7209	7384	407	15000
Precision	0.978	0.965		
Recall	0.940	0.950		
F-Measure	0.959	0.958		