On the Design of an Artificial Player for a Popular Word Game

Alberto Coffrini, Stefania Monica, and Federico Bergenti

ITALIAN CONFERENCE ON COMPUTATIONAL LOGIC (CILC 2021)

September 9^{th}, 2021

Introduction

The Task

- This presentation reports on a software system designed to accomplish a specific natural language processing task.
- The targeted task is a challenge of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian proposed in 2020.
- The task consinst in solving “La Ghigliottina”, the closing game of a popular Italian TV show.

An Example of the Game

Clues

- Corrente (Italian for *current*)
- Fuoco (Italian for *fire*)
- Passione (Italian for *passion*)
- Moda (Italian for *fashion*)
- Giorno (Italian for *day*)

Solution

Settimana (Italian for *week*)

Dataset Construction

Text file collection → Text cleaner → Data Structure

Text File Collection

- Books
- Corpus Paisà
  500 MB of texts from the Web
- Wikipedia articles
  (Around 150) from different topics
- Dump Wiki-Title-IT
  3,5 milion of titles
- Idiomatic phrases, proverbs, and compound words
  Dictionary Tullio de Mauro

Token and Related Token

Token and Related Token

- The process of finding solutions to the considered word game is based on finding relations between pairs of words.
- Pairs are extrapolated by parsing cleaned texts and by defining two subsequent words in the same sentence as a pair.
- The first word of each pair is denoted as *token* and the second word is denoted as *related token*.

$$< token, related\ token >$$

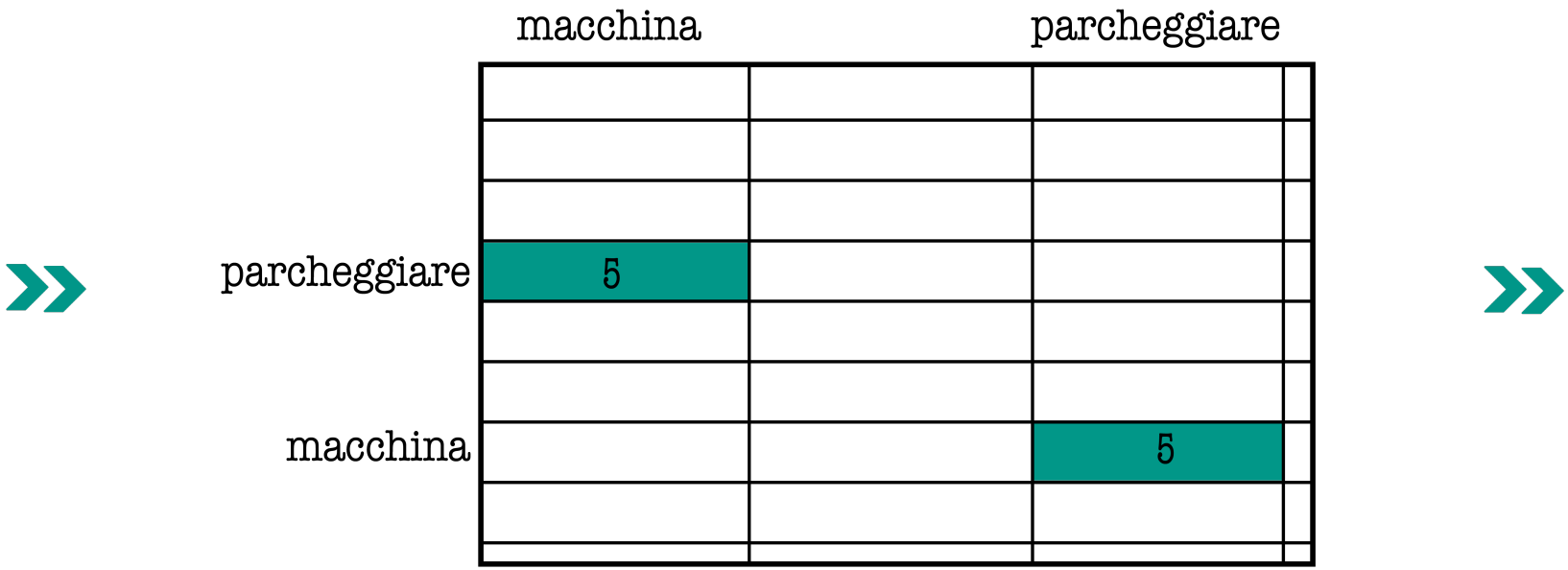
Example

$< parcheggiare, macchina >$
(Italian for $< parking, car >$)

Example

Occurence

The occurrence is equal to the number of times in which the pair is found in cleaned texts.



Match and Frequency

Match

The match of a generic related token r_j is the number of tokens for which r_j is a related token.

Frequency

The frequency of the generic related token r_j , denoted as f_j, is evaluated as the sum of the values $\{o_{i,j}\}_i^5 = 1$

$$f_j = \sum_{i=1}^{5} o_{ij}.$$

j: Token, i: Related Token.

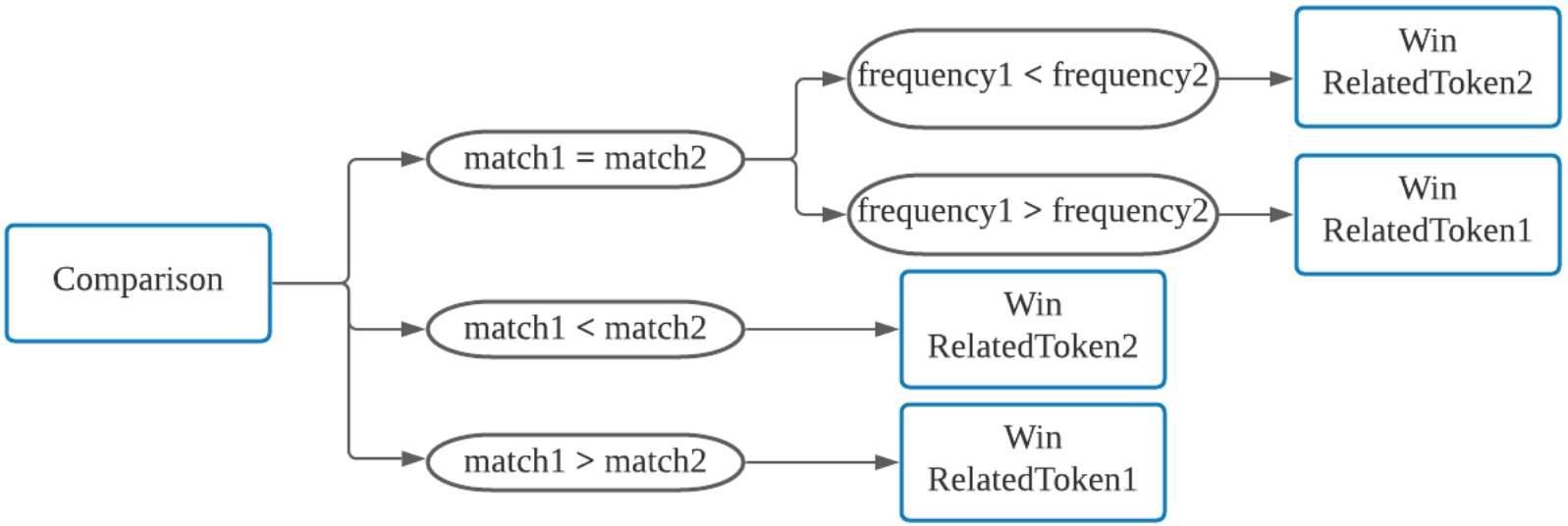
Token and Related Token

$$
\begin{array}{llll}
< t_1, r_1 >, & < t_1, r_2 >, & < t_1, r_3 >, & < t_1, r_4 > \\
< t_2, r_1 >, & < t_2, r_3 >, & < t_2, r_4 > & \\
< t_3, r_1 >, & < t_3, r_2 >, & < t_3, r_4 >, & \\
< t_4, r_1 >, & < t_4, r_3 >, & < t_4, r_4 >, & < t_4, r_5 > \\
< t_5, r_1 >, & < t_5, r_4 >, & < t_5, r_6 >, & < t_5, r_7 >, \quad < t_5, r_8 >
\end{array}
$$

$$
\begin{array}{llll}
o_{11} = 2, & o_{12} = 1, & o_{13} = 2, & o_{14} = 1 \\
o_{21} = 1, & o_{23} = 2, & o_{24} = 2 & \\
o_{31} = 5, & o_{32} = 4, & o_{34} = 8 & \\
o_{41} = 3, & o_{43} = 3, & o_{44} = 7, & o_{45} = 1 \\
o_{51} = 7, & o_{54} = 8, & o_{56} = 5, & o_{57} = 1, \quad o_{58} = 9
\end{array}
$$

Best Related Token

- First, the set of candidate solutions is restricted to the related tokens with the largest match value.
- Then, the solution is chosen as the one with the largest frequency value.



Best Related Token

Data obtained from the solver

- Chimico (Italian for *chemical*)
- Polizia (Italian for *police*)
- Segreto (Italian for *secret*)
- Cambio (Italian for *change*)
- Commercio (Italian for *trade*)

	Italiano	Legame	Agente	Laboratorio
Match	5	3	5	3
Frequency	17	5	278	14

Success Rate

To test the validity of the proposed approach, 100 different instances of the word game La Ghigliottina were considered. The instances were taken from those actually proposed in the TV program.

The system produced a total of **24** correct answers (**24%**).

Frequency Correction

A Problem

Frequency value formation can lead to an incorrect result.

Occurence and Frequency Values for the Related Token k

$$o_{k1} = 1, \quad o_{k2} = 1, \quad o_{k2} = 2, \quad o_{k4} = 160, \quad o_{k5} = 1,$$

$$f(k) = 165$$

Occurence and Frequency Values for the Related Token j

$$o_{j1} = 8, \quad o_{j2} = 9, \quad o_{j3} = 20, \quad o_{j4} = 10, \quad o_{j5} = 15,$$

$$f(j) = 62$$

Frequency Correction

The Solution

After various empirical tests, a *Threshold* equal to 13 has been set for the frequency.

Occurence and Frequency Values for the Related Token k with Threshold

$$o_{k1} = 1, \quad o_{k2} = 1, \quad o_{k2} = 2, \quad o_{k4} = 13, \quad o_{k5} = 1,$$

$$f(k) = 18$$

Occurence and Frequency Values for the Related Token j with Threshold

$$o_{j1} = 8, \quad o_{j2} = 9, \quad o_{j3} = 13, \quad o_{j4} = 10, \quad o_{j5} = 13,$$

$$f(j) = 53$$

Success Rate

To test the validity of the proposed approach, 100 different instances of the word game La Ghigliottina were considered. The instances were taken from those actually proposed in the TV program.

- Correct answer **47%**.
- Wrong answer with high scoring solution **15%**.
- Wrong answer but appropriate for the 5 words given **28%**.

Comparison with Other Systems

- “Il mago della Ghigliottina” 68,6%
- “Gul.le.ver” 26%.

Telegram Bot

Conclusion and Future Work

Conclusion

A software system capable of playing 'The Guillotine' has been presented.
The first version of the solver had a performance of 24%. After some improvements it was able to guess the solution at 47%.

Future Work

- Extention of the dataset.
- Use metrics to compare words.
- Detailed Analysis of the solver.

Thanks for your attention