Ghigliottin-AI @ EVALITA2020: Evaluating Artificial Players for the Language Game "La Ghigliottina"

Pierpaolo Basile
Dept. of Computer Science
University of Bari, Italy
pierpaolo.basile@uniba.it

Marco Lovetere
Ghigliottiniamo
marlove@gmail.com

Johanna Monti and **Antonio Pascucci**
UNIOR NLP Research Group
"L'Orientale" University of Naples, Italy
{jmonti, apascucci}@unior.it

Federico Sangati
UNIOR NLP Research Group
"L'Orientale" University of Naples, Italy
OIST Graduate University, Japan
federico.sangati@gmail.com

Lucia Siciliani
Dept. of Computer Science
University of Bari, Italy
lucia.siciliani@uniba.it

Abstract

English. Evaluating Artificial Players for the Language Game "La Ghigliottina" (Ghigliottin-AI) task is one of the tasks organized in the context of the 2020 EVALITA edition, a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language. Ghigliottin-AI participants are asked to build an artificial player able to solve "La Ghigliottina", namely the final game of an Italian TV show called "L'Eredità". The game involves a single player who is given a set of five words unrelated to each other, but related with a sixth word that represents the solution to the game. Fourteen teams registered to Ghigliottin-AI. Nevertheless, only two teams submitted their run. In order to evaluate the submitted systems, we rely on an API base methodology, via a Remote Evaluation Server (RES). In this report we describe the Ghigliottin-AI task, the data, the evaluation and we discuss results.

1 Background and Motivation

Language games draw their challenge and excitement from the richness and ambiguity of natural language, and therefore have attracted the attention of researchers in the fields of Artificial Intelligence and Natural Language Processing. For instance, IBM Watson is a system which successfully challenged human champions of "Jeopardy!", a game in which contestants are presented with clues in the form of answers, and must phrase their responses in the form of a question (Ferrucci et al., 2010; Molino et al., 2015). Another popular language game is solving crossword puzzles. The first experience reported in the literature is Proverb (Littman et al., 2002), that exploits large libraries of clues and solutions to past crossword puzzles. WebCrow is the first solver for Italian crosswords (Ernandes et al., 2008).

Following the first edition of the NLP4FUN task (Basile et al., 2018), proposed at EVALITA 2018, we propose a new edition of the task whose aim is to design a solver for "The Guillotine" (La Ghigliottina, in Italian) game. It is inspired by the final game of an Italian TV show called "L'Eredità". The game, broadcast by Italian national TV, involves a single player, who is given a set of five words - the clues - each linked in some way to a specific word that represents the unique solution of the game. Words are unrelated to each other, but each of them has a hidden association

with the solution. Once the clues are given, the player has one minute to find the solution. For example, given the five clues: *pie, bad, Adam, core, eye* the solution is *apple*, because: apple-pie is a kind of pie; bad apple is a way of referring to a trouble maker; Adam's apple is the prominent part of men's throat; apple core is the centre of the apple; apple of someone's eye is way of referring to someone's beloved person. This report is organized as follows: in Section 2 we describe the *Ghigliottin-AI* task. In Section 3 we present the dataset. The task evaluation is in Section 4. Results achieved by participants are shown in Section 5. Conclusions are in Section 6.

2 Task Description

Evaluating Artificial Players for the Language Game "La Ghigliottina" (Ghigliottin-AI) is one of the fourteen EVALITA 2020 tasks (Basile et al., 2020). Ghigliottin-AI participants are asked to build an artificial player able to solve "La Ghigliottina". They can take advantage of solutions adopted by previous systems (Semeraro et al., 2009; Basile et al., 2016; Sangati et al., 2018) and the availability of open repositories on the web.

3 Dataset

We provided a set of 300 games with their solution taken from the last editions of the TV game as training data. The training data was released in JSON format as shown in Figure 1. In this example, the first JSON shows the clues "posto" (literally *place*), "artificiale"(*artificial*), "lavaggio" (*washing*), "allenare" (literally *to train*) and "gallina" (*chicken*) and the solution "cervello" (*brain*): *non avere il cervello a posto* (*to be nutty*), *cervello artificiale* (*artificial brain*), *lavaggio del cervello* (*brainwashing*), *allenare il cervello* (*stretch the brain*) and *cervello da gallina* (*hare-brained*). In the second JSON we find "essere" (*to be*), "comparsa" (*appearance*), "x men", "ronaldo" and "mondiale" (*global*) and the solution "fenomeno" (*phenomenon*): *essere un fenomeno* (*be a phenomenon*), *comparsa di un fenomeno* (*apperance of a phenomenon*), *Fenomeno* is one of the X-men, *Fenomeno* was Ronaldo's nickname and *fenomeno mondiale* (*worldwide phenomenon*).

The test set consists in 350 games instances, provided by a Remote Evaluation Server (RES)

```
[
  {
      "w1": "posto",
      "w2": "artificiale",
      "w3": "lavaggio",
      "w4": "allenare",
      "w5": "gallina",
      "solution": "cervello"
  },
  {
      "w1": "essere",
      "w2": "comparsa",
      "w3": "x men",
      "w4": "ronaldo",
      "w5": "mondiale",
      "solution": "fenomeno"
  },
  ...
]
```

Figure 1: JSON format of the training set.

Ghigliottiniamo[1] at random intervals of time as a request with a single game challenge to registered systems. The RES allowed the systems to reply with a single solution to the game. *Ghigliottiniamo*[2] currently enables both humans and artificial systems to submit solutions to the TV game in real-time.

4 Task evaluation

In order to evaluate the AI systems, we rely on an API based methodology. During the evaluation period, at random intervals of time (over a period of 7 days), the RES submitted 350 game challenges to the registered systems. The systems had to reply back to the RES with a single solution to the game.

As evaluation measure, we adopt the standard accuracy score:

$$\frac{\text{solved games}}{\text{total games}} \quad (1)$$

As in the TV game, where players have one minute to provide the solution, the RES will discard system solutions received after 60 seconds from the submitted challenge.

[1] https://quiztime.net

[2] https://play.google.com/store/apps/details?id=io.quiztime.game

5 Results

Fourteen teams registered to the Ghigliottin-AI task. However, only two teams participated to the final test: *GUL.LE.VER* (De Francesco, 2020) and *Il Mago della Ghigliottina* (Sangati et al., 2020). GUiLlotine gLovE resolVER (GUL.LE.VER) is based on the Glove (Pennington et al., 2014) vector representation of the words on the basis of a large collected dataset, containing the Italian Wiktionary, Wikiquote, Wikipedia (only titles), the Italian Collocations Dictionary and other resources scraped on the web containing Italian multiword expressions, proverbs and songs titles. The Glove algorithm was chosen for its intrinsic power in capturing the co-occurrence correlation between two words that are not synonyms, due to the co-occurrence matrix that the algorithm builds before the training. The solution is searched in the vector space near the clues, obtaining a list of solution candidates. This list is descending reordered using a hybrid function composed by two parts: one part is based on the Pointwise Mutual Information; the other one is based on the weighted sum of the cosine similarity between the candidate solutions and the clues, in which the weight is the normalized IDF of the single clue in the corpus (solutions that are correlated with the rarest clues are more important than others). *Il Mago della Ghigliottina* is the same system submitted with the name of *UNIOR4NLP* in the *NLP4FUN* task in 2018 without any changes. The system is based on the observation that most cases clues and solution are connected because they form a multiword expression. In addition, clues are almost always nouns, verbs or adjectives, while solutions are nouns or adjectives. The system is based on a number of freely available corpora, such as: Paisà[3]; itWaC[4]; Wiki-IT-Titles downloaded via WikiExtractor[5]; 1955 proverbs from Wikiquote[6] and 371 from an online collection[7] downloaded on the 24th April 2018. Further lexical resources were developed from "Il Nuovo vocabolario di base della lingua italiana" and from the "De Mauro online dictionary". Technical details about *Il Mago della Ghigliottina* are available in (Sangati et al., 2018), submitted for the *NLP4FUN* task.

Table 1 shows the results of the two systems.

System	Correct	Total	Acc.
GUL.LE.VER	94	350	0.269
Il Mago della Ghigliottina	240	350	**0.686**
Combined (upper bound)	257	350	0.734

Table 1: Results

Both systems were able to provide a solution to all 350 games within a minute. The recorded time of the two systems ranges between 0.316 and 9.988 seconds. It is important to keep in mind that in addition to the response time, the recorded time includes the latency of the network and the time required for the instance to wake-up if it is set to go to sleep when idle. *Il Mago della Ghigliottina* is the system with the highest accuracy (about three solutions out of four correct), followed by *GUL.LE.VER* which on average is able to solve one game out of four.

We have computed the upper bound of the accuracy of the two systems on the test set when used in combination. The resulting accuracy is 73.4%, about 5 percentage points above the best performing system. This means that the two systems have some complementary and could be used in combination with some aggregating strategy.

6 Conclusions

In this report we presented *Ghigliottin-AI*, one of the EVALITA 2020 task. Despite fourteen teams subscribed to the task, just two of them submitted their system, namely *GUL.LE.VER* and *Il mago della Ghigliottina*. This latter achieved the best performances in terms of accuracy (68.6%), while *GUL.LE.VER* obtained 26.9% of accuracy.

Systems have been evaluated through an API methodology conducted by the Remote Evaluation Server (RES) (*Ghigliottiniamo*). To our knowledge, this is the first time that an API based system has been used on a NLP evaluation task. We believe this methodology has a strong advantage compared to a manual evaluation, as systems can be tested more systematically, fairly and continuously in time. We strongly hope that more

[3] https://www.corpusitaliano.it/

[4] https://wacky.sslmit.unibo.it/doku.php?id=corpora\#italian

[5] http://attardi.github.io/wikiextractor.

[6] https://it.wikiquote.org/wiki/Proverbi_italiani

[7] http://web.tiscali.it/proverbiitaliani

tasks will adopt this evaluation strategy in the future. The Ghigliottiniamo system currently enables both humans and artificial systems to submit solutions to the *Ghigliottina* when a new game is broadcasted on TV. This will allow us in the future to compare their results more systematically. The system remains open for new artificial systems to join the live competition[8].

References

Pierpaolo Basile, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2016. Solving a complex language game by using knowledge-based word associations discovery. *IEEE Transactions on Computational Intelligence and AI in Games*, 8(1):13–26.

Pierpaolo Basile, Marco de Gemmis, Lucia Siciliani, and Giovanni Semeraro. 2018. Overview of the evalita 2018 solving language games (nlp4fun) task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Nazareno De Francesco. 2020. Gul.le.ver, a glove based artificial player to solve the language game "la ghigliottina". In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*.

Marco Ernandes, Giovanni Angelini, and Marco Gori. 2008. A web-based agent challenges human experts on crosswords. *AI Magazine*, 29(1):77.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79.

Michael L Littman, Greg A Keim, and Noam Shazeer. 2002. A probabilistic approach to solving crossword puzzles. *Artificial Intelligence*, 134(1-2):23–55.

Piero Molino, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis, and Pierpaolo Basile. 2015. Playing with knowledge: A virtual player for "who wants to be a millionaire?" that leverages question answering techniques. *Artificial Intelligence*, 222:157–181.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Federico Sangati, Antonio Pascucci, and Johanna Monti. 2018. Exploiting multiword expressions to solve "la ghigliottina". In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263, pages 258–263. Accademia University Press.

Federico Sangati, Antonio Pascucci, and Johanna Monti. 2020. "il mago della ghigliottina"@ghigliottin-ai when linguistics meets artificial intelligence. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*.

Giovanni Semeraro, Pasquale Lops, Pierpaolo Basile, and Marco De Gemmis. 2009. On the tip of my thought: Playing the guillotine game. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1543–1548, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[8] `https://quiztime.net`