NLP4FUN at EVALITA 2018
Task Guidelines

Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, Lucia Siliciani
University of Bari Aldo Moro, Italy
{nlp4fun.evalita@gmail.com

May 28, 2018

Contents

1 Task Description

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 [3, 5]. Another popular language game is solving crossword puzzles. The first experience reported in the literature is Proverb [4], that exploits large libraries of clues and solutions to past crossword puzzles. WebCrow is the first solver for Italian crosswords [2].

The proposed task consists in designing 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: *sin*, *Newton*, *doctor*, *New York*, *bad*, the solution is **apple**, because: the apple is the symbol of original sin in Christian theology; Newton discovered the gravity by means of an apple; "an apple a day keeps the doctor away" is a famous proverb; New York city is also called "the big apple"; and "one bad apple can spoil the whole bunch" is a popular phrase which figuratively means that the person doing wrong can have a negative influence on those around him. "La Ghigliottina" is a challenging language game which demands knowledge covering a broad range of topics. Artificial players for that game can take advantage from the availability of open repositories on the web, such as Wikipedia, that provide the system with the cultural and linguistic background needed to understand clues [1]. Participants must build an artificial player able to solve "La Ghigliottina".

2 Development and Test Data

2.1 Data Format

We provide a set of both training and testing games in the XML format:

```
<games>
  <game>
        <id>3fc953bd-bd48-4fb9-a86c-bd979c1b5c3f</id>
        <clue>uomo</clue>
        <clue>cane</clue>
        <clue>musica</clue>
        <clue>casa</clue>
        <clue>pietra</clue>
        <solution>chiesa</solution>
  </game>
  ...
</games>
```

The XML file consists of a root element *games* which contains several *game* elements. Each game has five *clue* elements and one *solution*. We provide 316 and 105 games as training and

testing, respectively. The participants **can integrate any knowledge resources in their systems except further games**.

3 Submission Format

Participants must provide for each game a ranked list of **maximum 100 tentative solutions**. Results must be provided in a single text plain file according to the following format:

```
id solution score rank time
```

Values must be separated by a whitespace character and time must be reported in milliseconds. For example:

```
3fc953bd-bd48-4fb9-a86c-bd979c1b5c3f porta 0.978 1 3459
3fc953bd-bd48-4fb9-a86c-bd979c1b5c3f chiesa 0.932 2 3251
3fc953bd-bd48-4fb9-a86c-bd979c1b5c3f santo 0.897 3 4321
...
3fc953bd-bd48-4fb9-a86c-bd979c1b5c3f carta 0.321 100 2343
...
```

4 Evaluation

As evaluation measure, we adopt a weighted version of Mean Reciprocal Rank (MRR). Since time is a critical factor in this game, the Reciprocal Rank will be weighted by a function which takes into account the time. In the TV game, the player has one minute to provide the solution. Taking into account these factors, the final evaluation measure is:

$$\frac{1}{|G|}\sum_{g\in G}\frac{1}{r_g}max(\frac{1}{t_g},\frac{1}{10}) \tag{1}$$

where G is the set of games and r_g is the rank of the solution, while t_g denotes the minutes taken by the system to produce the tentative solutions. **Systems that take more than 10 minutes are equally penalized**.

5 How to submit your runs

The test data will be distributed on September 10th, 2018. Once you have run your system over the test data, you will have to send your results to us following these recommendations:

- each team can submit maximum two results files
- choose a team name and name the file containing your runs in the following way: $nlp4fun2018.teamName.run1$ $nlp4fun2018.teamName.run2$
- send the file to the email address: nlp4fun.evalita2018@gmail.com using the subject: $nlp4fun - teamName$

References

[1] P. Basile, M. de Gemmis, P. Lops, and G. Semeraro. 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, 2016.

[2] M. Ernandes, G. Angelini, and M. Gori. A web-based agent challenges human experts on crosswords. *AI Magazine*, 29(1):77, 2008.

[3] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

[4] M. L. Littman, G. A. Keim, and N. Shazeer. A probabilistic approach to solving crossword puzzles. *Artificial Intelligence*, 134(1-2):23–55, 2002.

[5] P. Molino, P. Lops, G. Semeraro, M. de Gemmis, and P. Basile. Playing with knowledge: A virtual player for "who wants to be a millionaire?" that leverages question answering techniques. *Artificial Intelligence*, 222:157–181, 2015.