Natural Language Generation system writing football articles

Author: Dan Raffl | Supervisor: RNDr. Jiří Hana, Ph.D. | 2023

Introduction

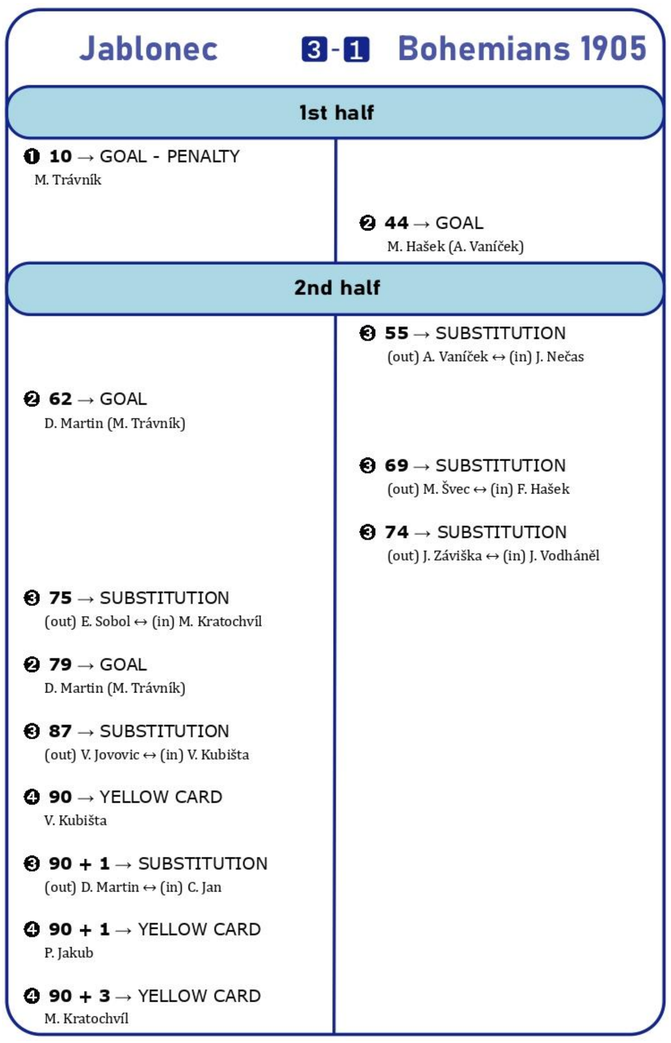
The goal was to develop a rule-based Natural Language Generation (NLG) system automatically producing articles in Czech on football matches based on structured data.

The main focus was on content selection, document planning and microplanning.

Realisation is delegated to another system. This thesis presents such an end-to-end system implemented in Python along with a brief overview of NLG process.

Results

Input



Output

Jablonec porazil Bohemians 1905 3:1.

V desáté minutě proměnil Michal Trávník penaltu. Hašek dal po přihrávce Vaníčka Antonína 44 minut po začátku branku. 55 Minut po začátku druhého poločasu vystřídal hráč s číslem 10 Vaníčka. Po nádherné souhře v 62. minutě vstřelil Martin Doležal po asistenci Trávník Michal gól. 69 Minut po začátku druhého poločasu Hašek střídal za hráče s číslem 2. V 74. minutě vystřídal Jan Vodháněl za hráče s číslem 8. Miloš Kratochvíl střídal za Sobola Eduarda 75 minut po začátku. 79 Minut po začátku druhého poločasu vsítil Martin Doležal po pase Trávník branku. 87 Minut po začátku Kubišta Vojtěch vystřídal Vladimir Jovovič. Devadesát minut po začátku druhého poločasu obdržel Kubišta Vojtěch žlutou. V první minutě nastavení druhého poločasu střídal Chramosta za Martina Doležala. Povazanec dostal jednu minutu po začátku nastaveného času druhého poločasu žlutou kartu. V třetí minutě nastavení druého poločasu vyfasoval Kratochvíl Miloš žlutou.

Approach

The system exploits modular architecture creating a one-way pipeline to connect the modules. Here are the core modules and their brief description:

1) Data initializer

Transforms data from the initial JSON file to a more convenient inner representation implemented as a system of classes.

2) Document planner

Decides the content of the article. Output of this module is a set of non-verbal messages and their order in the outputted text.

3) Sentence planner

Module combining lexicalization and referring expression generation tasks. The core process is to select a template for each non-verbal message. Template consists of immutable lexical tokens and entities. Entities are then suitably lexicalized under constraints regarding repetition. The result is a text with grammatical tags (for Genja).

4) Linguistic realiser

Realisation is performed by Genja API, therefore this module handles the communication with the API (forming the valid input, calling the API, processing the output).

Conclusions

The quality of text is not on the sports journalist level, however we have created an end-to-end NLG system that takes raw data as input and writes Czech text summarizing the football match. To improve the quality of the article we propose to use different approach or to broaden the dataset information (e.g., goal types, stats of the game, players knowledge database, etc.).

Further information

1. Realisation system: Genja API (Contact Geneea at geneea.com)
2. For the complete thesis please see https://github.com/Mokeas/BcThesis

1 2

Acknowledgments

I would like to thank my supervisor, RNDr. Jiří Hana, Ph.D., for providing the dataset and the system for realisation as well as his guidance.