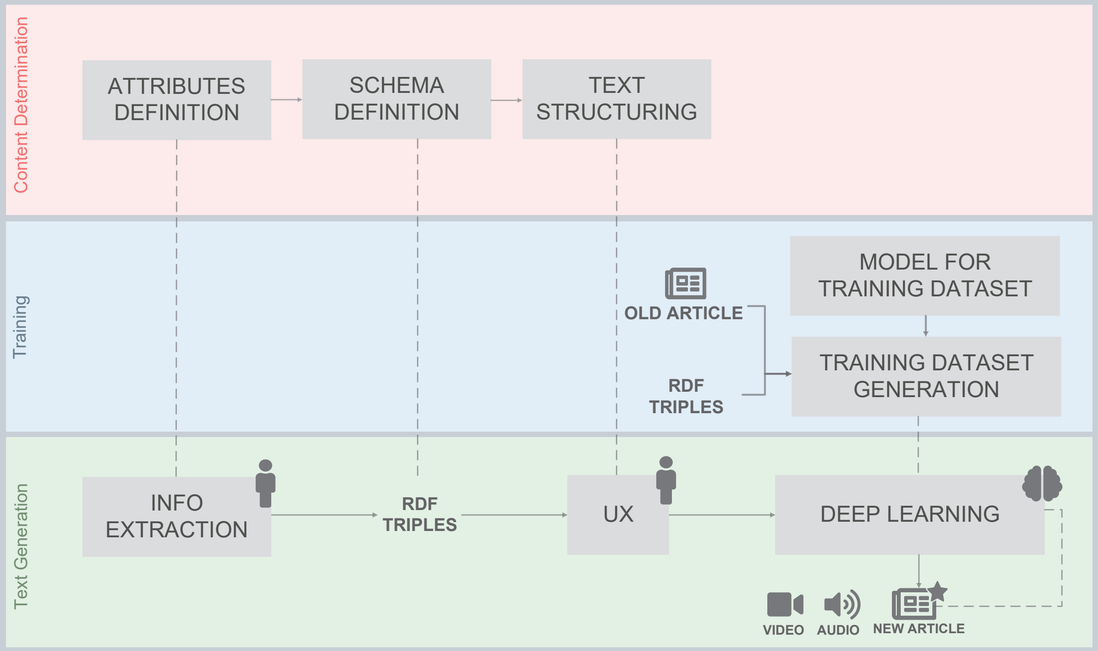
Supporting **Journalism** by combining **Natural Language Generation** and **Knowledge Graphs**

Marco Cremaschi, *Federico Bianchi, Andrea Maurino, Andrea Primo Pierotti*
marco.cremaschi@unimib.it

University of Milano-Bicocca, Milano, Italy



Gazette Lexicalization

GazelLex is a new prototype that implements a **Neural Machine Translation** (*NMT*) approach through the use of deep learning techniques to generate soccer articles (sentences) starting from data gathered from an online provider and converted in **RDF triples**.
GazelLex is also able to generate **videos** containing the images and the prominent information of the article, and to generate **audio** using a speech synthesis module.

1 CONTENT DETERMINATION

To select the most relevant information, GazelLex uses an handcrafted approach. One of the primary references was **PASS***, a personalised automated text system developed to write soccer articles. The selected data are used to create triples for the following phases.

* van der Lee et al., 2017

Example of Entities

TEAM · FORMATION · COACH

Example of Predicates

injuryAt · yellowCardAt · violentFoulAt

2 TEXT STRUCTURING

Being a domain specific process, GazelLex uses some handcrafted developed **template** based on real articles.

It is possible to find pre-made templates (e.g. *complete* or *short article*) but it is also possible to modify them or create new ones.

3 SENTENCE AGGREGATION

RDF triples are aggregated to generate sentences with less redundancy to make the article more concise and coherent.

4 NEURAL LEXICALIZATION

RDF data are converted into natural language using **LSTM***.

- no limitations in input and output length
- input and output are not independent

The neural architecture is based on a standard encoder-decoder structure with 4 LSTM layers containing 200 hidden neurons on both the encoder and the decoder. Input tokenization is based on the space character.

* Hochreiter and Schmidhuber, 1997

5 REFERENCE EXPRESSION GENERATION

Different databases are used to avoid redundancy and give a fluent text to the reader

- **DBpedia** (list of possible replacements for a team or players' name)
- **Wikidata** (list of soccer teams nicknames)
- **Topend Sports database**