Supporting Journalism by Combining Neural Language Generation and Knowledge Graphs

Marco Cremaschi, Federico Bianchi, Andrea Maurino, Andrea Primo Pierotti
Department of Computer Sciences, Systems and Communications
University of Milan-Bicocca
Viale Sarca, 336 - 20126, Milan, Italy
{marco.cremaschi,federico.bianchi,andrea.maurino}@unimib.it
a.pierotti1@campus.unimib.it

Abstract

Natural Language Generation is a field that is becoming relevant in several domains, including journalism. Natural Language Generation techniques can be of great help to journalists, allowing a substantial reduction in the time required to complete repetitive tasks. In this position paper, we enforce the idea that automated tools can reduce the effort required to journalist when writing articles; at the same time we introduce GazelLex (Gazette Lexicalization), a prototype that covers several steps of Natural Language Generation, in order to create soccer articles automatically, using data from Knowledge Graphs, leaving journalists the possibility of refining and editing articles with additional information. We shall present our first results and current limits of the approach, and we shall also describe some lessons learned that might be useful to readers that want to explore this field.

1 Introduction

Although automation is a phenomenon that is becoming more and more visible today, there are specialised jobs that require human effort to be completed. The job of a journalist is among these (Örnebring, 2010). However, recent technological progress in the field of Natural Language Generation (NLG) and the use of increasingly sophisticated techniques of artificial intelligence allow the use of software capable of writing newspaper articles almost indistinguishable from human ones. These techniques can help journalists reduce the effort needed for repetitive tasks, such as data collection and drafting writing. The name given to this phenomenon is Automated Journalism; this new type of journalism uses algorithms to generate news under human supervision. During the past years, several newsrooms have begun to experiment this technology: Associated Press, Forbes, Los Angeles Times, and ProPublica are among the first, but adoption could spread out soon (Graefe, 2016). Automated Journalism can bring a massive change to the sector: writing news is a business that endeavours to minimise costs while maintaining maximum efficiency and full speed, and thanks to this software the above-mentioned objectives can be achieved, generating good-quality articles (van Dalen, 2012). This new technology provides many advantages: the most evident are speed and the scale of news coverage. Of course, there are also problems and limitations. One of the most relevant is the dependence from structured data (Graefe, 2016), that is the reason why sports reports, financial articles, and forecasts are the most covered topics by software: they are all domains where the complexity of the topic can be managed from software using structured data. Similar structured data are not always available in other fields. In order to generate valuable text, approaches considering data contained in the Knowledge Graphs (KGs) have recently been introduced in literature (Gardent et al., 2017; Trisedya et al., 2018).

A Knowledge Graph (KG) describes real-world entities and the relations between them. KGs are an essential source of information, and their features allow the use of this information in different contexts, such as link prediction (Trouillon et al., 2016) and recommendation (Zhang et al., 2016). Popular KGs are the Google Knowledge Graph, Wikidata and DBpedia (Auer et al., 2007). Entities are defined in an ontology and thus can be classified using a series of types. The primary element of a KG to store entities information is a

Resource Description Framework (RDF) triple in the format $\langle subject, predicate, object\rangle$. As RDF triples open many possibilities in Web data representation, utilising this data also in the NLG context is valuable (Perera et al., 2016). Interlinked KGs can be used to automatically extend the information relating to a given entity in an article.

In our solution, we use DBpedia, one of the fastest growing Linked Data resource that is available free of charge; it is characterised by a high number of links from the Linked Data Cloud[2]. DBpedia is thus a central interlinking hub, an access point for retrieving information to be inserted in an article, as specified below.

Up to 2010, commercial providers in the NLG field were not popular, but in the last years few companies have started to provide this kind of services. In 2016 there were 13 companies covering this field (Drr, 2016) (e.g., AutomatedInsights[3], NarrativeScience[4]). Approaches that try to integrate deep networks and text generation are now common in literature (Gardent et al., 2017). These automated tools are going to become a standard method to help journalist during the news writing process.

We shall concentrate on examples of related work in the context of lexicalization from RDF data, we shall refer to surveys from the state of the art for a more detailed overview of the field (Reiter and Dale, 1997; Gatt and Krahmer, 2018; Moussallem et al., 2018). Semantic web technologies like RDF can be used to enhance the power of current algorithms (Bouayad-Agha et al., 2012). The WebNLG challenge (Gardent et al., 2017) has been introduced to study the possibilities given by the combination of deep learning techniques and semantic web technologies. In a similar context, an approach based on Long Short-Term Memory (LSTM) networks has been proposed to generate text lexicalizations from RDF triples (Trisedya et al., 2018).

In this work, we aim to describe what is the possible automation process that can be used to help journalist in the news writing process. At the same time we describe a new prototype we have created to support journalistic activities, GazelLex (Gazette Lexicalization). GazelLex, through the use of deep learning techniques implements a Neural Machine Translation (NMT) approach to generate articles (sentences) starting from data composed by 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 (Figure 1). To the best of our knowledge, our prototype is the first to provide an all-in-one integrated approach to NLG with RDF triples in the context of helping journalist in writing articles.

This paper is structured as follows: in Section 2, we analyse the state-of-the-art on Natural Language Generation, showing that these methods to generate natural language are becoming popular. In Section 3 we describe our prototype, GazelLex, that combines neural methods and knowledge graphs to create soccer articles and describe how this kind of tools can be of help to journalism. In Section 4 we show a preliminary experimental analysis, while in Section 5 we provide conclusions.

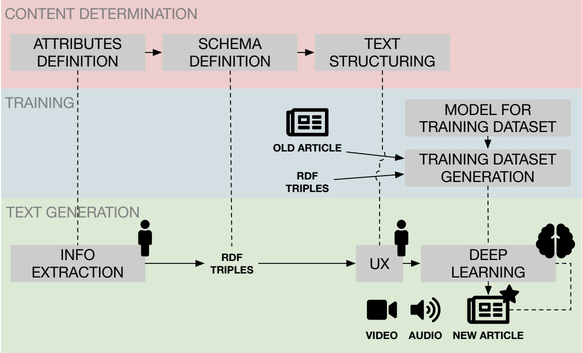


Figure 1: The workflow of our model.

2 Natural Language Generation

NLG is a "sub-field of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information" (Reiter and Dale, 1997; Reiter and Dale, 2000). In NLG six "problems" must be addressed: **Content determination**: input data that is always more detailed and richer than what we want to cover in the text (Gatt and Krahmer, 2018) and so the aim is to filter and choose what to say. **Text structuring**: a clear text structure and the order of presentation of information are critical for readers, for this reason, pre-defining the templates is necessary. **Sen-**

[2] https://wiki.dbpedia.org/dbpedia-2016-04-statistics

[3] https://automatedinsights.com/

[4] https://narrativescience.com/

tence aggregation: sentences must not be disconnected. Text needs therefore to be grouped in such a way that a “more fluid and readable” text (Gatt and Krahmer, 2018) is generated. **Lexicalization**: one of the most critical phases of NLG process is how to express message blocks through words and phrases. This task is called lexicalization and concerns the actual conversion from messages to natural language. **Reference expression generation**: to avoid repetitions, selecting ways to refer to entities using different methods (such as pronouns, proper nouns, or descriptions) is essential. **Linguistic realisation**: it concerns the combination of relevant words and phrases to form a sentence.

As we stated above, lexicalization is one of the most critical and complex tasks in the NLG process. Natural language vagueness and choosing the right words to express a concept are intricate issues to manage. Looking at the state-of-the-art, we see that recent research on this topic shows that an interesting solution in these cases is based on Machine Learning (ML) (Gatt and Krahmer, 2018). Moreover, a recent challenge in the NLG field, launched and published in 2017, called WebNLG (Gardent et al., 2017) confirms the idea that not only we need to combine ML methods to generate language, but we can also use KGs to enrich sentences with additional contextual information (e.g., contextual information about a player).

3 GazelLex

In this section, we shall give an example of the NLG process in a domain specific view. As introduced previously, we developed a software, named GazelLex, that can produce soccer articles. There are two main reasons for this choice: first of all, the project was partly commissioned by an Italian newspaper publisher. Furthermore, soccer and sports, in general, are good domains to develop NLG, because they are complex enough to be challenging, yet they are easy to manage and many data exist (Barzilay and Lapata, 2005). In this scenario we focused our attention on the final output, using a solution that combines neural network with some handcrafted processes. We would like to underline that the data related to the games (e.g. number of goals, training) are extracted automatically from online services.

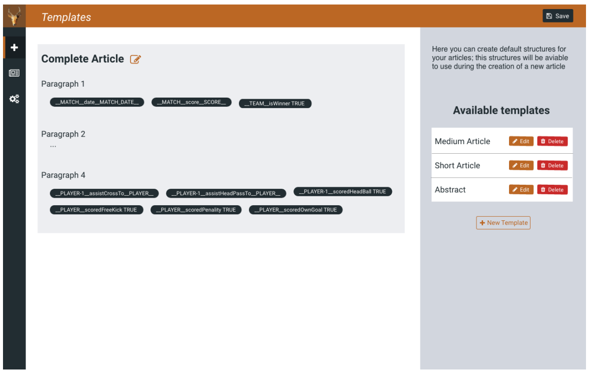
Our approach is divided into five tasks, in order to address the five classic NLG sub-problems (Gatt and Krahmer, 2018): in the following, for each phase, implementation details will be provided.

3.1 Content Determination

To select the most relevant information, a handcrafted approach was chosen. To select the information to bring in the final output, we traced the most used data in soccer articles. One of the primary references was PASS, a personalised automated text system developed to write soccer articles (van der Lee et al., 2017). We took the kind of information PASS used to fill its templates and enriched them with our data fields. So we have some entities of type “TEAM”, “FORMATION”, “COACH” and some predicates like “injuryAt”, “yellowCardAt”, and “violentFoulAt”[5]. The software used this data to create triples, that algorithms used to write the article.

3.2 Text Structuring

Being a domain specific process, we developed a handcrafted template, based on real articles. Aiming to get a similar output we imitated the journalist’s job in the division of text and about information contained in each part. We also considered the text structuring approach usually developed in this domain, that uses more general information and after that a chronological order (Gatt and Krahmer, 2018). In GazelLex, it is possible to find templates (e.g., complete or short article) resulting from the process described above, but it is also possible to modify them or create new ones (Fig. 2).



Complete Article

Paragraph 1

__MATCH__date__MATCH_DATE__ __MATCH__score__SCORE__ __TEAM__isWinner TRUE

Paragraph 2

...

Paragraph 4

__PLAYER-1__assistCrossTo__PLAYER__ __PLAYER-1__assistHeadPassTo__PLAYER__ __PLAYER-1__scoredHeadBall TRUE __PLAYER__scoredFreeKick TRUE __PLAYER__scoredPenality TRUE __PLAYER__scoredOwnGoal TRUE

Here you can create default structures for your articles; this structures will be aviable to use during the creation of a new article

Available templates

Medium Article

Short Article

Abstract

Figure 2: Page for creating and editing templates related to Text Structuring.

[5] A list of all the entities and the attributes are available here: https://goo.gl/LonnQ5

3.3 Sentence aggregation

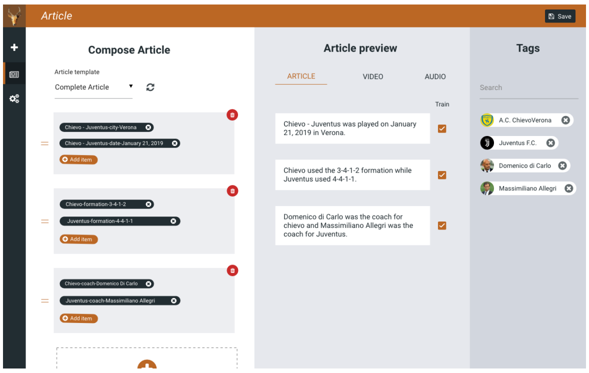
In soccer data, many events could be redundant when written in an article. If a player scores a hat trick in a match writing the same sentence about each goal would be unpleasant to read while grouping them in a single sentence could be more concise and coherent. This task "focused on domain- and application-specific rules" (Gatt and Krahmer, 2018). We aggregated the RDF triples defined in the preceding section to generate a group of triples that represents the content of our news article.

3.4 Neural Lexicalization

Like we said above, we considered lexicalization like a NMT process, converting RDF data into natural language. To achieve this aim, we used a specific kind of neural network: LSTM (Hochreiter and Schmidhuber, 1997). Their recent success in NLG field is related to many advantages they provide. Compared to the traditional neural network, LSTM do not have limitations in input and output length. Furthermore, input and output are not independent, that is a vital advantage in language generation. To predict a word in a sentence it is useful to know and consider the previous one, and the hidden states of the network keep the memory about what happened in previous timesteps. In this way, LSTM can combine the previous state, the memory collected and the input, allowing dependencies to be maintained in the long term. We experimented NMT using a now widely recognized tool for neural machine translation[6] (Klein et al., 2017). Our 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 (recall that our RDF triples' elements are separated by spaces).

3.5 Reference expression generation

We used different databases to avoid redundancy and give a fluent text to the reader. Some online resources help us to create a list of possible replacements for a team or players' name. Using DBpedia, we can find a nickname for an entity (Real Madrid players are also called *Blancos* or *Merengues*). Other resources we used are Wikidata list of soccer teams nicknames and Topend Sports database.



Article template: Complete Article

Chievo - Juventus-city-Verona
Chievo - Juventus-date-January 21, 2019

Chievo-formation-3-4-1-2
Juventus-formation-4-4-1-1

Chievo-coach-Domenico Di Carlo
Juventus-coach-Massimiliano Allegri

Chievo - Juventus was played on January 21, 2019 in Verona. ☑

Chievo used the 3-4-1-2 formation while Juventus used 4-4-1-1. ☑

Domenico di Carlo was the coach for chievo and Massimiliano Allegri was the coach for Juventus. ☑

A.C. ChievoVerona
Juventus F.C.
Domenico di Carlo
Massimiliano Allegri

Figure 3: Page for the revision of the lexicalized triples.

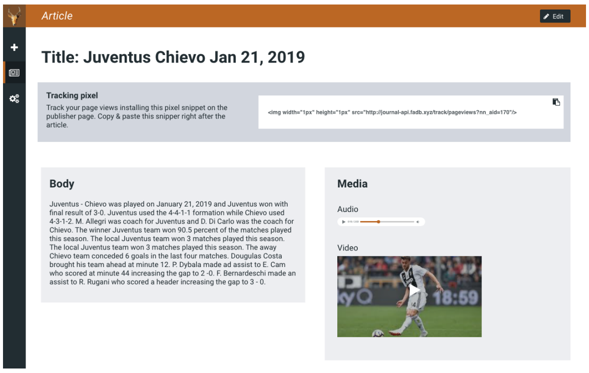
4 Analysis

In the following section, we shall show some insights into our tool and on how it works. We shall present a use case, a recent soccer match, for which the generation process and the resulting text will be shown. The initial dataset for the training was created manually and consists of 4387 pairs of triples and lexicalizations. We drew inspiration from the state-of-the-art to devise the architecture of our network (Gardent et al., 2017; Trisedya et al., 2018). From our primary experiments the best performing model required two layers of bidirectional LSTM, but still, the model suffers from some limitations (outlined in the related sec.).

4.1 Use Case Exploration

To show the valid output of GazelLex, we took an example match and generated its lexicalization. We considered the football match played by Juventus F.C. and A.C Chievo on the 21st of January. Our application gathered data from an online provider and converted data in a triple format. A journalist can edit settings using a form (Figure 3): the journalist is in charge of deciding what is worth writing in the article and how it should appear to the end-user; we recall that we can also define templates for our articles (Figure 2). The final output of this process looks like the one that is shown in Figure 4. GazelLex, in order to improve the quality of the sentences and to obtain results as close to the style of the journalist as possible (i.e. style transfer), cyclically re-executes the training phase using the sentences validated by the journalist. The following is an example of lexicalization of triples relative to the use case (Table 1).

[6] http://opennmt.net



Title: Juventus Chievo Jan 21, 2019

Tracking pixel

Track your page views installing this pixel snippet on the publisher page. Copy & paste this snipper right after the article.

```
<img width="1px" height="1px" src="http://journal-api.fadb.xyz/track/pageviews?nn_aid=170"/>
```

Body

Juventus - Chievo was played on January 21, 2019 and Juventus won with final result of 3-0. Juventus used the 4-4-1-1 formation while Chievo used 4-3-1-2. M. Allegri was coach for Juventus and D. Di Carlo was the coach for Chievo. The winner Juventus team won 90.5 percent of the matches played this season. The local Juventus team won 3 matches played this season. The local Juventus team won 3 matches played this season. The away Chievo team conceded 6 goals in the last four matches. Dougulas Costa brought his team ahead at minute 12. P. Dybala made ad assist to E. Cam who scored at minute 44 increasing the gap to 2 -0. F. Bernardeschi made an assist to R. Rugani who scored a header increasing the gap to 3 - 0.

Figure 4: Lexicalization of triples from the Juventus-Chivevo football match.

Table 1: Example of lexicalization.

$\langle Paulo\ Dybala, assistTo, Emre\ Can\rangle$ $\langle Emre\ Can, scoredAt, 44\rangle$ $\langle Emre\ Can, scoredWithScore, 0-2\rangle$
Paulo Dybala made an assist to Emre Can who scored at minute 44 increasing the gap to 0 - 2.

4.2 Current Limitations and Lessons Learned

In this section we would like to outline the current limitations of our project and also report a few lessons learned that might be useful for other researchers who are currently exploring this field. One key part of the development process comes from the definition or the selection of a good Knowledge Graph that can support the lexicalization; moreover, the definition of the new RDF predicates is a difficult process that must be done carefully to avoid errors in the next steps. Our application currently supports the lexicalization of a small set of triples (i.e., we focused on goals and final result); we decided to concentrate on this small set to generate a set of resulting sentences that can be manually inspected for quality. Our NLG model is based on a deep learning architecture, and thus some of the generated sentences are not well-formed owing to the structure of the net itself. While this is a problem that has to be solved in our settings, we have a journalist reviewing the article before it is released to the public: this allows us to have a model that is more flexible than standard pattern-based NLG, while the precision of the output can be controlled in a human-in-the-loop setting. Regarding the configuration of our model, we have replicated the state-of-the-art experiments (i.e. approaches explained in (Gardent et al., 2017)) and we are currently experimenting those architectures on our domain dataset. The results are yet to be quantitatively validated and they are preliminary, but they are promising as reported by journalists. In the future, we are planning to carefully explore various architecture and consider the use of word embeddings to solve some of our current issues.

5 Conclusion

In this position paper we have analysed the future possibilities given by automated journalism. We have summarised the current state of art on this topic showing that there is an increasing interest towards automated natural language generation for the news sector. While hereby, we showed an application related to the soccer domain, the principles and the methodologies described are general, and they can be used in other fields (e.g., finance, weather reporting). We strongly believe that these tools can greatly help journalists in working on what is really important (e.g., investigation, fact checking), leaving high effort, but low value tasks to computers. The prototype we have described is a first step towards this automated process and its results are surely promising.

Acknowledgments

This research has been supported in part by the Robo-Journalism project 2018-comm25-0047 in collaboration with Instal S.r.l.[7]. Special thanks to Carlo Mattioli and Alessandra Siano for their support during the development of the project.

References

[Auer et al.2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Barzilay and Lapata2005] Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 331–338, Stroudsburg, PA, USA. Association for Computational Linguistics.

[7] instal.com

[Bouayad-Agha et al.2012] N Bouayad-Agha, G Casamayor, and L Wanner. 2012. Natural language generation and semantic web technologies. *Semantic Web Journal*.

[Drr2016] Konstantin Nicholas Drr. 2016. Mapping the field of algorithmic journalism. *Digital Journalism*, 4(6):700–722.

[Gardent et al.2017] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133. Association for Computational Linguistics.

[Gatt and Krahmer2018] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170, January.

[Graefe2016] Andreas Graefe. 2016. Guide to automated journalism. Technical report, Tow Center for Digital Journalism, Columbia University.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[Klein et al.2017] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

[Moussallem et al.2018] Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1 – 19.

[Örnebring2010] Henrik Örnebring. 2010. Technology and journalism-as-labour: Historical perspectives. *Journalism*, 11(1):57–74.

[Perera et al.2016] Rivindu Perera, Parma Nand, and Gisela Klette. 2016. Realtext-lex: A lexicalization framework for rdf triples. *The Prague Bulletin of Mathematical Linguistics*, 106(1):45 – 68.

[Reiter and Dale1997] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87, March.

[Reiter and Dale2000] Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

[Trisedya et al.2018] Bayu Distiawan Trisedya, Jianzhong Qi, Rui Zhang, and Wei Wang. 2018. Gtr-lstm: A triple encoder for sentence generation from rdf data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1627–1637.

[Trouillon et al.2016] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

[van Dalen2012] Arjen van Dalen. 2012. The algorithms behind the headlines. *Journalism Practice*, 6(5-6):648–658.

[van der Lee et al.2017] Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104. Association for Computational Linguistics.

[Zhang et al.2016] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362. ACM.