# The Value and Benefits of Data-to-Text Technologies

David Martínez de Lecea[1], Darío Arenas[1], Javier Pérez[1], David Llorente[1], Eugenio Fernández[1], Antonio Moratilla[1]
[1](*Computer Science; University of Alcalá; Spain)*

***ABSTRACT:*** *Data-to-text technologies present an enormous and exciting opportunity to help audiences understand some of the insights present in today's vasts and growing amounts of electronic data. In this article we analyze the potential value and benefits of these solutions as well as their risks and limitations for a wider penetration. These technologies already bring substantial advantages of cost, time, accuracy and clarity versus other traditional approaches or format. On the other hand, there are still important limitations that restrict the broad applicability of these solutions, most importantly in the limited quality of their output. However we find that the current state of development is sufficient for the application of these solution across many domains and use cases and recommend businesses of all sectors to consider how to deploy them to enhance the value they are currently getting from their data. As the availability of data keeps growing exponentially and natural language generation technology keeps improving, we expect data-to-text solutions to take a much more bigger role in the production of automated content across many different domains.*

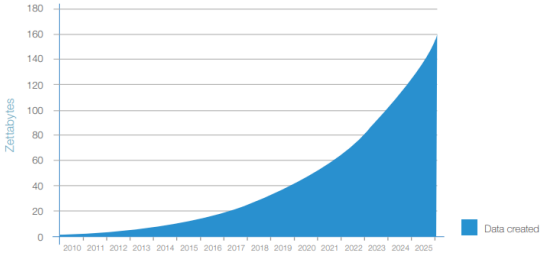
***KEYWORDS*** *- artificial intelligence, data-to-text, natural language, natural language generation*

## I. INTRODUCTION

The world generated ~30 zettabytes ($10^{21}$) of data in 2017, according to IDC [1]. And the clear consensus expectation is for this number to keep increasing at an exponential rate as the number of connected devices keeps growing (thanks to technologies such as the Internet of Things) and the cost of electronic data storage keeps decreasing, see figure 1.

A lot of data, which can contain very beneficial knowledge, remains unused and unanalyzed due to the human limitations to process such vasts amounts of information in a practical manner.

Fig. 1. Annual size of the global datasphere




Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

Natural Language Generation (NLG) is a subarea of the field of Natural Language Processing (NLP) that focuses on the construction of understandable text in natural language automatically. Data-to-text (D2T), a key component of NLG, covers the transformation of data into natural language. The key difference between data-to-text versus other areas of NLG is that the input information does not come in other form of natural language but in the shape of non-linguistic electronic data. This data is frequently structured data, i.e. stored as fixed fields within a record or file such as relational databases and spreadsheets, but, thanks to recent and spectacular development of deep learning technologies, it could also be unstructured data such as images, audio or video records whose content can be translated to text and subsequently used to produce long narratives with data-to-text techniques.

## II. DATA-TO-TEXT TECHNIQUES AND APPLICATIONS

### II.1. OVERVIEW OF TECHNIQUES

Even though most companies use proprietary technology for their data-to-text solutions, there seems to a be a common approach

that can be found in the academic literature o in open source tools such as SimpleNLG. In their 2017 paper "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation" [2], Gatt and Krahmer present a very comprehensive overview of the different approaches used to address data-to-text problems. What follows in this section is largely a summary of their work.

Many authors split the NLG problem of converting input data into output text, into a number of subproblems, most frequently these six:

1. **Content selection**: deciding which information will be included in the output text. In most situations there is more information in the data than what we want to include in the output text. This steps involves choosing what to include and what not and tends to be domain specific.

2. **Text structuring**: determine in which order information will be presented. In this step the data-to-text system determines the order in which the information selected will be presented. This step also tends to be domain specific but recently some researchers such as Barzilay & Lee [3] or Lapata [4] have explored the use of machine learning techniques to perform this step automatically.

3. **Sentence aggregation**: decide which information to present in each individual sentence. If each individual message is in its own independent sentence, the output text lacks flow and feels very artificial. This steps combines different messages into sentences. Once again, it is common to use domain-specific solutions but there are some efforts to use more generic approaches such as the work done by Harbusch & Kempen with syntactic aggregation to eliminate redundancy [5] [6].

4. **Lexicalisation**: find the right words and phrases to express the desired information. This is the step where the messages start to be converted into natural language. In many domains the goal is not only to generate correct language but also to produce a certain amount of variation.

5. **Referring expression generation**: in the words of Reiter and Dale [7]: "*the task of selecting words or phrases to identify domain entities*". The same authors differentiate this against lexicalization stating that referring expression generation is a "*discrimination task, where the system needs to communicate sufficient information to distinguish one domain entity from other domain entities*" [8].

6. **Linguistic realisation**: combine all words and phrases into well-formed grammatically and syntactically correct sentences. This task step requires ordering the components of a sentence, and generating the right morphological forms (e.g. verb conjugations and agreement), function words and punctuation marks. The main approaches for this include human-crafted templates, human-crafted grammar-based systems and statistical techniques. It will be obvious that the former produces very accurate but inflexible output whereas the latter can generalize to many more domains and use cases as the expense of assured correction.

The presented six tasks are normally organized in one of the following three ways:

1. **Modular architectures**: Frequent in traditional systems that address the NLG problem through the symbol-processing paradigm that early AI research favored. These architectures use clear divisions among sub-tasks. The most common (also known as the 'consensus') architecture of these systems was described by Reiter in 1994 [9] and includes three modules:

The *text planner* (which combines the tasks of *content selection* and *text structuring*) is concerned with "what to say" and it produces a *text plan*, with a structured representation of the messages, which is used as the input of the next module. This step tends to be combined with the work of domain experts to build a knowledge base so that the information is stored in a way that captures semantic relationships within the data.

The *sentence planner* (which combines the tasks of *sentence aggregation*, *lexicalisation* and *referring expression generation* [8]) decides "how to say it" and produces a *sentence plan.*

The *realizer* (*linguistic realization*) finally produces generates the final *output text* in a grammatically and syntactically correct way by applying language rules.

2. **Planning perspectives**: These systems view text generation as planning. In the field of AI, planning is described as the process of identifying a sequence of actions to achieve a particular goal. This paradigm is used in NLG by viewing text generation as the execution of of actions to satisfy a

communicative goal. This approach produces a more integrated design without such a clear separation between tasks. There are two main approaches within this perspective.

*Planning through the grammar* views linguistic structures as planning operators. This approach requires strong grammar formalism and clear rules.

*Stochastic planning under uncertainty using Reinforcement Learning*. In this approach, text generation is modelled as a Markov decision process: states are associated with possible actions and each state-action pair is associated with a probability of moving from a state at time t to a new state at t + 1 through action a. Transitions are associated with a reward function that quantifies the optimality of the output generated. Learning is usually done simulating policies –different paths through the state space– that are associated with a reward that measures its optimality. This approach could also be considered to belong to next category: data-driven approaches.

3. **Data-driven, integrated approaches**: Having seen the great results of machine learning solutions in other areas of AI, strong effect described as "The Unreasonable Effectiveness of Data" by Halevy, Norvig and Pereira at Google [10], the current trend in NLG is to rely on statistical machine learning of correspondences between non-linguistic inputs and outputs. These approaches also render more integrated approaches than those of the traditional architectures. There are different approaches within this category depending on whether they are based on language models, classification algorithms or seen as inverted parsing. Analyzing those approaches is beyond the scope of this article but it is worth mentioning that there is one area that is gathering most of the attention of the research community these days: deep learning methods.

Fueled by the success of deep neural architectures in other both related (machine translation) and unrelated areas of AI (computer vision or speech recognition), there have been a number of efforts to apply these techniques to natural language processing and generation. The architecture that seems to be showing most promising results in natural language is that of long short-term memory (LSTM) recurrent neural networks (RNN). These structures include memory cells and multiplicative gates that control how information is retained or forgotten, which enables them to handle long-range dependencies commonly found in natural language texts. In 2011, Sutskever, Martens and Hinton, used an LSTM RNN to generate grammatical English sentences [11], and, since then, many other NLG applications of deep neural networks have been tried. In particular, there have been some promising results in applying neural networks to data-to-text generation such as Mei, Bansal and Walter's application to weather information [12] or Lebret, Grangier and Aulli's to Wikipedia biographies [13].

## II.2. APPLICATIONS

The potential applications of data-to-text are innumerable. Any regular report or communication based on structured input data is that a business produces is a candidate to be automated. The following is non-exhaustive selection as examples.

**Journalism**. Generate automatic news articles. Data-heavy topics such as sports, weather of financial markets are very well suited for these applications. In fact, this is one of the areas with the current highest penetration of data-to-text solutions.

**Business intelligence tools**. In order to extract meaning, numbers require calculations, graphs require interpretation. NLG enhances the, normally graphical, output of business intelligence tools for analysis and research. Users can then benefit from a combination of graphs and natural language to bring to life the insights found in the data.

**E-commerce**. Generate the descriptions of products being sold online. This is particularly useful for marketplace businesses that can offer millions of different products with high SKU rotation and in multiple languages. These product descriptions could even be generated taking account the particular needs of interests of each individual customer greatly enhancing the effectiveness of the, already very successful recommendation systems of the big E-commerce players.

**Business management**. Generate required business reports for management, customers or regulators. This saves a lot of time or highly paid staff so that they can focus in decision making, not on report writing. These benefits are not only economic but also bring more professional

satisfaction to those users who can spend more of their time on higher value-adding tasks.

**Customer personalization**. Create personalized customer communications. Reports can be generated for "an audience of one". These can even be generated in real time when users interact with the business through digital channels providing them with the most up-to-date information.

## III. BENEFITS, LIMITATIONS AND RISKS

### III.1. BENEFITS

The use of data-to-text technologies present some very clear benefits over more traditional approaches, namely:

**Cost**. Data-to-text technologies can produce content at a cost at least 10 times lower than the traditional alternative of human writing, even higher in domains with highly-skilled, highly-paid staff. This even permits unearthing content that otherwise would never be analysed and commented on due to lack of resources or the impracticality of writing for very small audiences. Data-to-text makes feasible writing narratives for an audience as small as one person.

**Production time**. The conversion of data into text, once the system has been setup, is done in a an instant. This is of particular interest in areas where it is important to report on recent information e.g. news, or where a business is providing instant feedback on a customer's situation, e.g. personalized report on financial situation generated in real-time.

**Accuracy**. Computers do not make mistakes or write typos. As long as the data source is correct, the output text can be always correct without any potentially embarrassing and distracting mistakes. On the other hand, whenever there are issues with the input data, data-to-text solutions tend not to be able to identify them and produce output than can be seriously misleading.

**Clarity**. Information presented in natural language is easier to understand. In an experiment from 2017, the researchers found that the use of Natural Language Generation content enhances decision-making under uncertainty, compared to state-of-the-art graphical-based representation methods. In a task-based study with 442 adults, they found that presentations using NLG led to 24% better decision-making, on average, than the graphical presentations, and to 44% better decision-making when NLG is combined with graphics. They also found this effect to be potentially stronger in women who achieved an 87% increase in results quality, on average, when using NLG compared to just graphical presentations [14].

**Scale**. As long as there's new data, there is no theoretical limitation as to how much output volume can be generated. Hence, once the solution has been deployed, its benefits can be enjoyed by an unlimited number of users.

### III.2. LIMITATIONS

However, the current technological development of these solutions present some limitations to the broad applicability of these solutions.

**Lack of creativity**. Firstly, the current output of data-to-text systems still presents some repeatability in its format. Machines cannot yet produce as much variety of content as humans can. Contrary to the belief of many, this is not due to the inability of computers to show create new ideas but to the limitations imposed by the data, which takes us to the second point.

**Limited scope**. Currently, machines using data-to-text techniques can only talk about the information present in the dataset used as input. This drastically compares with the human ability to bring information from other sources such as common knowledge of the field or to use analogies from other fields.

**Setup time**. In order to deploy data-to-text technologies to new areas, a significant development effort is required to customize the general techniques to the specific domain of interest.

**Rigidity**. Current state-of-the-art techniques still require important customization of the deployments for each specific use case. Whenever there are important changes in the data source or the application needs, the solutions need to be retrained and reconfigured.

### III.3. RISKS

The most prevalent risk in the application of these solutions comes from its limitations in the scope of their analysis work. Whatever is not in the data just does not exist for these tools. Should there be an important outside event impacting the

domain of interest, the data-to-text system would not be able to incorporate it into its output which might generate extremely short-sighted and useless material. Should someone rely on that content to make decisions, there would a significant risk of wrong decisions being made.

Any analysis of automation technologies would not be complete without considering the risks that they pose to the jobs of people performing the tasks that are getting automated.

At the moment, the application of data-to-text technologies produces a clear net benefit to those that use them. This benefit normally comes in two shapes. Some are using these solutions to generate content that otherwise they would not be generating. For example a sports news site can use data-to-text technologies to cover minor competitions that otherwise would be to expensive to cover. Others leverage these technologies to expedite their production of content by automating the creation of first drafts, for example. "*The immediate opportunity isn't to fully automate the research process but to make it more structured and efficient.*" says Brian Ulicny, data scientist at Thomson Reuters Labs [15].

Although these cases do not seem to imply any potential negative implications in the very short term, there are some considerations for their impact in the longer term. By using these solutions to write content of lesser importance, they might be replacing the work previously done by apprentices and new entrants to the profession, which might create hinder the ability of these people to develop their skill set. This might require re-thinking the apprenticeship model for the development of junior staff in several industries.

Finally, in an unclear but possible scenario, it is plausible that these techniques will keep improving and at some point overcome the limitations exposed, which will greatly exacerbate the risks and implications of too much automation.

## IV. BARRIERS TO THE EXPANSION OF DATA-TO-TEXT

Despite the attractiveness of data-to-text solutions as presented in the previous section, the implementation of these remains still quite limited. This can be explained by the following factors.

**Lack of awareness**. Most people are not aware of the existence of data-to-text solutions and are very surprised the first time they are presented with them and the quality of their output. In simple terms, these technologies are not used more often because potential users do not know they exist.

**Lack of input data**. Not all knowledge domains have enough data of the required volume and quality to produce interesting content through data-to-text. Frequently the data just does not exist or the quality is not sufficient for a data-to-text system without a very significant effort to preprocess it.

Even when the data exists, it's sometimes owned by one particular entity that does not have the knowledge, the expertise or the interest to exploit it in this fashion foregoing the opportunity to extract insights and value from that information.

**Fear of substitution**. When potential users come across applicable data-to-text solutions, these tend to be the very same people whose content generation work might get replaced, which puts them in a difficult position to properly and objectively assess the quality of the output of the tools and might tend to dismiss them.

**Lack of service providers**. As per our research and discussions with specialists in the field of Natural Language Generation, we have identified fewer than 10 companies providing generic cross-industry data-to-text services to third parties in the world. Taking into consideration the work required to deploy these solutions into each new domain, the world needs a much bigger number of data-to-text specialists if these technologies are going to become widespread.

## V. CONCLUSION

In this article, we have analyzed the potential value and benefits of data-to-text technologies as well as their risks and limitations for a broader penetration across many domains.

Their clear advantages versus the time and costs that would be incurred in producing the same content manually are large and obvious. However, the current state-of-the-art technology in this field presents important limitations that restrict the broad applicability of these solutions.

We find that, given the current state of development of these solutions, there is a great opportunity to expand their application to new domains and that those opportunities will only keep increasing, which presents great future prospects for the technology and those involved with it. Hence we recommend businesses of all sectors to consider applying these technologies to increase the value they are currently getting from their data.

As the availability of data grows and NLG technology keeps developing and improving, we expect data-to-text solutions to take a much more important role in the production of natural language text content across many different domains and applications.

## REFERENCES

[1] David Reinsel, John Gantz and John Rydning; Data Age 2025: The Evolution of Data to Life-Critical; *IDC, sponsored by Seagate, April 2017*

[2] Albert Gatt, Emiel Krahmer; Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation; *arXiv 1703.09902v1, March 2017*

[3] Regina Barzilay, Lillian Lee; Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. *NAACL, pp. 113–120, 2004*

[4] Mirella Lapata; Automatic Evaluation of Information Ordering: Kendall's Tau; *Computational Linguistics, 32 (4), 471–484, 2006*

[5] Karin Harbusch, Gerard Kempen; Generating clausal coordinate ellipsis multilingually: A uniform approach based on post editing. *ENLG, pp. 138–145, 2009*

[6] Gerard Kempen; Clausal coordination and coordinate ellipsis in a model of the speaker. Linguistics, 47 (3), 653–696, 2009

[7] Ehud Reiter, Robert Dale; Building Natural Language Generation Systems; *Building natural-language generation systems. Natural Language Engineering, 3, 57–87, 1997*

[8] Ehud Reiter, Robert Dale; Building Natural Language Generation Systems; *Cambridge University Press, Cambridge, UK, 2000*

[9] Ehud Reiter; Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?; *International Workshop on Natural Language Generation, pp. 163–170, 1994.*

[10] Alon Halevy, Peter Norvig, Fernando Pereira; The Unreasonable Effectiveness of Data; *IEEE Intelligent Systems, March/April 2009*

[11] Ilya Sutskever, James Martens, Geoffrey Hinton. Generating Text with Recurrent Neural Networks. *Proceedings of the 28th International Conference on Machine Learning (ICML), pp. 1017–1024, 2011*

[12] Hongyuan Mei, Mohit Bansal, Matthew R. Walter. What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment; *Proceedings NAACL-HLT'16, pp. 1–11,* 2016

[13] Remi Lebret, David Grangier, Michael Auli. Neural Text Generation from Structured Data with Application to the Biography Domain; *Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016*

[14] Dimitra Gkatzia, Oliver Lemon, Verena Rieser; Data-to-Text Generation Improves Decision-Making Under Uncertainty; *IEEE Computational Intelligence Magazine, volume 12, issue 3, August 2017*

[15] Walter Frick; Why AI Can't Write This Article (Yet); *Harvard Business Review, 24 July 2017*