Assessment 2

Natural Language Generation

CS551H

Yassin Dinana

52094480

12th May 2021

Introduction:

Natural Language Generation, better known as NLG, is the process of transforming structured data to human readable text as an output. It is used in many different fields nowadays as organizations can fully depend on the NLG technology to generate reports and papers. The most important feature when building an NLG system is the data, if the data is enough and correctly pre-processed the generated output will be high quality [1].

There are different approaches used when building an NLG system, those approaches are rule-based and machine learning techniques, where the rule-based approach follows a template where the text is pre-written, it also follows a set of stages of planning the data before generating the output text. As the technology emerged, new approaches have been introduced such as the Machine Learning and Deep Learning approaches, those are used for document structuring, content selection, and optimal ordering. Different NLG tasks will also be discussed in this report where it covers different technologies used over time [2].

This assessment report will be discussing different NLG approaches with a deep technological explanation as well as comparing the old technologies with the new ones based on different attributes such as the cost, quality, user-friendliness and control, and the integration of different technologies.

This report will be assessing each method on Aberdeen's City Council as a use-case context where NLG solutions are used on dashboards highlighting and reporting social activities happening in Aberdeen.

Lowering Development Cost:

The NLG market is expected to grow by 16.1% from $35.6 Million to $1.161.3 Million by 2027. Developing an NLG system is costly; therefore, it is important to calculate the cost needed to build an NLG system. Different approaches will have different costs; therefore, this section will be evaluation the cost of different approaches and then compare them. This section will also include the cost based on old and new technology and evaluate if new technology will affect the cost [3].

To calculate the development cost, different areas are included in the cost calculations which will be found in the tables below.

Parameter	**Cost**	**Details**
Salaries (Per Year)	£180.000	This cost is based on salaries where there are five engineers: - **1 junior engineer:** £40.000 - **1 mid-level engineer:** £50.000 - **1 senior engineer:** £60.000 - **1 Marketing Specialist:** £30. 000
Computers Prices	£21600	This is the cost for 12 iMac Apple computers where each is £1800

Cloud and Servers	£12000	This price is for starting companies which is a subscription at Amazon Web Services cloud to store data – the £1000 per month offers a deal for companies.
Marketing and Sales	£17.500	This is based on the subscriptions price for the software to be $350 for each user and the aim is to reach 1000 users by the first year – the marketing cost is 5% of the predicted revenue which is £17.500.
Rent	£12000	This is based on Aberdeen's rent prices – An apartment office can be £1000 per month – that is £12000 per year.
Energy and Utilities	£5400	This is based on Aberdeen's prices for electricity and utilities – where electricity is ranged from £300 per month and water £150 per month for the water.

Looking at the table above, the total cost for the first year to develop an NLG software is £248500 – Where the predicted revenue as stated in the marketing and sales section of the table will be £350. 000, therefore, the profit is £101,500 before tax which is 29% [3].

The number of employees and cost above is based on the rule-based approach, which is the basic approach when building an NLG system and is considered old technology as it includes the basic steps to developing the system. New technology has been introduced to the system in the past few years such as the Machine Learning approach. When researching to use machine learning for building the NLG system, it is concluded that it will be costly at the beginning because of the following reasons:

- Buying and acquiring more data to cope with ML.
- Buy more powerful computers and processors to handle training
- Buy more cloud space for new data

Therefore, it is concluded that using new technology can be more costly at the beginning, but it will lower the cost on the long run; this is because the new technologies such as ML require stronger software and hardware components, but it is faster in processing, meaning that it will save time and therefore save cost as this might lead to less human power on the long run.

Maintaining High Output-Text Quality:

After calculating all the required cost for the system to be built, it is now important to maintain the high output-text quality when deploying the system. There are different requirements needed to make sure the output is of high quality, the output quality is not only based on how the system is built

but it is also based on the input data that are fed to the system when it is deployed, the requirements are:

- *Input Data:* The most important aspect to make sure the output generated is of high-quality is to make sure that the data is enough and correctly pre-processed.
- *Content Determination*: Determine which information from the data are important to be used and structuring all the information. This is also known as Macro Planning.
- *Sentence Planning*: This process decides how the data and important information must be split into difference sentences and paragraphs, this is known as Micro Planning, the process includes different techniques such as lexicalization and grammaticalization.
- *Surface Realizer*: This covers the selection of the important syntax and inflection.
- *Physical Presentation*: This area will also be covered in the integration section of this report in more detail, it covers how the output should be presented, it can be presented as speech such as Siri and Alexa or written text, it covers the punctuation and articulation.

Following those steps will guarantee a high-quality, but when dealing with NLG, high-quality text might come with constraints, a constraint that is met a lot in NLG is the vocabulary constraint, where all words generated as an output must follow the specific vocabulary rules from language generated. The NLG approach that has the most complaints is the Neural NLG, particularly sequency-to-sequence models, constraints that are usually met when dealing with challenging datasets include:

- Mapping from conceptual input to lexical items
- Punctuation markers
- Noun Phrases
- Lexical Choice

Rule-Based approach deals easier with all constraints as the steps for building the system are being followed, when using newer but more complicated approaches such as Neural NLG will result in more constraints such as the ones stated above, it is also harder to fix the constrains using the newer ML approaches.

When dealing with the dashboard for Aberdeen's City Council, large reports and text will be generated, therefore, it is advised to start by using the Rule-Based approach as it produces less issues in the text and if any, they are easier to fix [5].

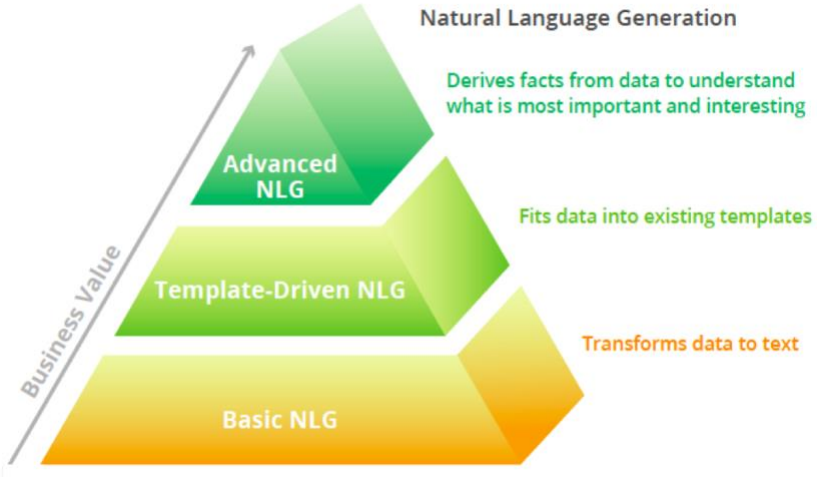

Figure 1: NLG technologies-based businesses

Offering user-configurable NLG solutions:

As the developed NLG system will be used by different users all around the world, the system needs to be configurable and user-friendly for the users.

1- The system must be able to generate text in different languages – this will require more data in all languages that the developer wishes to include, and it will lead to more cost. However, it will allow the NLG software to reach more people speaking different languages which will increase the subscriptions and reaching more users which is the main goal.

2- When the system generates an output for a user, the output should not always be paragraphs, some sentences can be generated short and easy which will allow the users to easily understand the generated text, and this will also decrease the possibility of any mistakes in the generated sentences.

Using new technology will ease the generation of user-configurable NLG solutions, using machine learning to train all the new data given as an input as well as all the data collected while users are experiencing the system. New technology plays a big role in such area as some NLP and NLG applications collect data in real-time when users are using the system, this will help understand users from different backgrounds and ease the use of the system for them. This is done by using Machine Learning and Neural Networks to:

- Gather data in real-time
- Generate Output
- Save output in case there are any bugs to fix later

Visual NLG is an NLG platform that is very user friendly and generates easy output in different languages, it supports users at all levels using their cloud-based services. It is advised that to achieve user-configurable NLG solutions is to use advanced technology on the long run-in order to increase the size of the business and reach new users in Aberdeen.

Integrating into adjacent technologies such as speech, data analytics, and dashboards:

Both methodologies follow a data-to-text approach to generate the desired document. However, so far, we have discussed documents that are purely textual. Since Aberdeen City Council intends to use the perspective NLG system for storytelling; integrating such a system with existing technologies that enhance the storytelling, such as adding figures and images, can be very powerful in communicating ideas.

Neural system receives data input into the network and produce the output text. Developers have very little control over how the text presents itself. This is mainly due to the black box nature of neural networks; we cannot infer what's which nodes inside the network affect the output most. This aspect of neural NLG systems makes it difficult to integrate them with services such as Tableau or Microsoft Power BI.

For Rule-Based NLG systems, the transparent nature of the modelling enables the developers to have direct control over the parameters of the system. Consequently, integrating such systems with services such as the ones is a significantly easier task than is the case with neural systems.

For example, market leader in NLG systems is Arria NLG which offers their cutting edge Arria studio licence for a fee. The studio can easily be integrated with a variety of services including the ones mentioned above.

Conclusion:

Through this technology assessment report, we have outlined the pros and cons of using either machine learning NLG or rule-based NLG systems. We summarized estimated costs for Aberdeen council for adopting either approach. Then we highlighted the ease of use, integration with current technologies and the quality of output of each approach.
It is our recommendation that Aberdeen City Council uses Rule-Bases NLG system for the services required for the first five years from 2022 to 2027, followed by a gradual shift toward utilising modern machine learning NLG approaches. Our justification for this recommendation relies on two pillars:

1- Using machine learning in NLG is a new technology that has not been fully realised yet. This is one reason why many existing companies are hesitant to change their systems from rule-based to neural ones now. However, as with any new technology, this situation will change. The technology will improve with time and will become cheaper to utilise and deploy and eventually it will take overrule-based systems. Therefore, Aberdeen Council would be in a good position when the time comes.

2- The initial costs of setting up a neural system are exceedingly high particularly for a publicly funded institution. Those costs are not justified for the purpose and for the quality they yield. Thus, kickstarting Aberdeen NLG services with rule-based systems will as 1) training wheel for current and future talent in the Council. 2) Create some revenue to finance further expansion into neural systems in the future.

References:

1- V. Duc. T. Son. “Natural Language Generation for Non-Expert Users”. Online. Available at: https://arxiv.org/pdf/1909.08250.pdf [Accessed 13/5/2021]
2- M. Klarner. “A Scalable Natural Language Generation Approach”. Online. Available at: https://www.aclweb.org/anthology/W04-2809.pdf [Accessed 11/5/2021]
3- “Reports and Data”. Online. Available at: https://www.reportsanddata.com/report-detail/natural-language-generation-nlg-market [Accessed 11/5/2021]
4- A. Gatt. E.Krahmer. “Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation”. Online. Available at: https://research.tilburguniversity.edu/en/publications/survey-of-the-state-of-the-art-in-natural-language-generation-cor [Accessed 9/5/2021]
5- Devepedia. Online. Available at: https://devopedia.org/natural-language-generation [Accessed 9/5/2021]