Module 2 - Automated Content Production

Hello and welcome back for week two.

Today we're going to talk about automated content production, but before we get into today's lecture I want to remind you that there's another video lecture this week as well as a video demo of an automated writing tool that you should also watch.

Also there are three readings that I've posted in the syllabus this week including an excerpt from the guide to automated journalism, a piece on how automation is changing financial news coverage, and a case study of how automation has been used to cover local elections in Switzerland. Also, don't forget to engage in the forums this week and to take the quiz.

So this week's topic again is all about automated content production, and today in particular I want to give you an overview of some of the use cases and scenarios where we've already seen news organizations using automation to generate publishable news content.

So let's jump right into some examples. Here we see an example excerpt of an automatically written article produced using some polling data from the 2016 U.S. elections. It comes from www.pollyvote.com which is a site you can check out online if you're interested.

Now, oftentimes what we're talking about when it comes to automated content is a system of written templates. Those templates are usually written by people and those templates have a certain placeholders which can be substituted in with words based on data and based on logical If/Then/Else rules that the template writer writes into the template.

You can see in this example where different things have been filled in from data based on the underlying style that's shown. In some cases there are placeholders marked where various synonyms might be inserted and which can then provide some variability in the text.

In other cases you see placeholders where data is taken directly from the raw data and inserted into the template or other cases where data is derived from the input data to produce the text to fill in.

This is another example of an article that was generated automatically. The Associated Press produces these articles every fiscal quarter. In fact it produces thousands of articles this way in order to bolster its financial earnings coverage.

You can see that the articles are pretty straightforward, maybe even a bit dry, but they do get the main facts across, so it's got the profit, the earnings per share, the revenue forecasts, and so on.

In some cases these automatically generated articles are actually augmented by human reporters who will do additional reporting and then add context into the article. So in those cases, it's more of a hybrid

process where the automation produces the initial first version like the one you're seeing here and then a human adds in some additional context and insight in a second pass.

I also want to draw your attention to the end of this story in this example where it provides information about where the data came from and what software was used to create the content. This Is sort of a byline for automated content, but I like it because it provides some transparency for how the content was created.

Here's another example of an automatically produced article. This one coming from the Washington Post and its U.S. election coverage in 2016. They had article pages like this for each of the state elections for instance.

Now one thing to point out here is that in 2012 the Washington Post only covered 15 percent of the congressional races in the U.S. In 2016, they covered 100 percent of the federal elections including all the House Senate and gubernatorial elections in the United States, and these articles were also dynamic over the entire course of election night.

Yet another example is this site called Cluck Spark from Sweden. What this site does is it's a sports site and it provides automatically generated articles that have been produced for all of these soccer or football games that have been played throughout Sweden.

So there's six divisions of soccer in Sweden, and what this site does is it publishes an article for all of those matches that occur. So every local game will have an article written about it.

Typically they're short articles maybe 100 words, and again they're very factual and sort of straightforward based on the scores and the events that happened in the game. It's not really too fancy. Any given story might recount who scored the goals as well as maybe a little bit of history or some league standing for the teams that are playing, but what's interesting is that the automation provides a foundational breath for coverage.

So anyone looking for quick facts about a local match of soccer played in Sweden will be able to find that story on the site. In addition to the automatically generated articles there are 14 sports reporters who work with the site and who can look at the automatically generated stories to find ones that might be more interesting or might be interesting to do additional reporting on, and then reporters can go out and do that additional reporting, enhance the articles or in fact write entirely new articles on top of the automatically written versions of the articles.

Now, all of the examples that I've just shown you are examples of automated content that are produced using the standard NLG model. NLG stands for natural language generation, and it's the technology that's used to produce the automated writing outputs in these various projects.

So the standard NLG model consists of three phases: document planning, micro planning, and realization.

Document planning has to do with determining what to communicate in the story and then how to structure it. So this has to do with figuring out what's interesting, what's newsworthy, what should I include in my story, and then it figures out how to structure those things and order those things into a narrative or into a descriptive explanation of what's going on in the event.

The next phase is called micro planning. This has to do with more word and sentence and phrase level decisions. So which word am I going to use in this sentence, which which phrase am I going to use, what kinds of synonyms do I want to include, and so on.

And then finally there's the realization phase and this has mostly to do with grammar rules making sure that the text that's generated is grammatically correct. So things like verb conjugation or noun fertilization need to be attended to so that the text that's ultimately output by the system is correct. There are also a number of other approaches and some interesting opportunities for automated content production.

So again all of the examples that I've shown you so far in this presentation have been using this standard NLG approach, but there are also statistical techniques that are out there that can be used to generate text.

These techniques don't use templates, instead they're trained on lots and lots of data. So examples of articles and a machine learning techniques can be used to then learn models from those examples and then generate new texts based on the models that are learned.

Typically statistical techniques are not as good as template based techniques in terms of the quality of text that's produced, and so we don't see a lot of these statistical techniques actually being used in industry right now.

Another approach to text production relates to summarization. So the process of taking a longer article and compressing it down into a short summary version of that article. There's different approaches here. There's abstractive approaches and extractive approaches to summarization. An extractive approach for example, if you imagine an article with 10 sentences, an extractive approach would maybe extract two of those sentence, and it would just copy those two sentences out of the article and then it would blew them together, and those two sentences would become the summary.

An abstractive summarization algorithm, on the other hand, actually has a potential to generate new texts, new sentences, that don't exist in the original article. So it can actually get a lot more sophisticated in the text that it can produce. Producing new words, new phrases, and new ordering of information in the summary.

Again because of the quality of these algorithms currently abstractive techniques don't tend to be used as much as extractive techniques. There's also some exciting opportunities in terms of automatically producing other forms of media.

So we could talk about non textual output media like video. You can try tools like Wibbitz or Watch it, which are semi automated tools for generating video. You can also produce data visualizations

automatically and in some cases we see examples of bots using automated content generation to produce texts that then gets shared on social media. So here's a quick example of an automatically generated data visualization.

[00:10:33] This is actually comes from Spiegel Online. It is sort of a past diagram for a soccer game and you can see that it shows the different names of players in red, and then sort of who passed to who as arrows.

What's interesting about these diagrams is that typically they're not published automatically. They're generated automatically, but they're not published automatically, and the reason for that is that the sports reporters like to use these as a basis for their own storytelling and for their own interpretation. So they'll go in and look at these automatically generated diagrams and then do some of their own interpretation and write a story that explains the overall strategy that the teams took in the game.

Another example here again I mentioned bots in this example the Treasury iRobot summarizes U.S. treasury spending that it collects every day, it generates a tweet length written description, and then it tweets it out.

So one other thing I'd like to talk about today is about how automation changes work. Automated content will always need people to be involved in and assessing and reassessing how can be approved. There are new tasks for people involved in this, and roles will also evolve as the technology advances.

So given the prevalence of the template driven approaches to automated content in practice I would say that writing is one of the areas where people will definitely need to evolve their craft. Template writers will need to approach a story with an understanding of what the data could potentially say and to be able to think about you know all the ways that the data could give rise to different angles and different stories, and then be able to articulate the logic that would drive those various variations in a template.

People will also need to be involved in monitoring and in fixing errors that crop up in automated content output. This could involve debugging the system as well as more routine maintenance activities updating, tweaking, and editing things. When a data stream is updated, knowledge bases or databases may need to be updated to reflect those changes.

If new data sources are released or updated or their format has changed people will need to be involved with updating the templates and rules for reading that data. There's also, I think, going to be some evolution of different roles in the newsroom.

So, as newsrooms expand their use of automation, people will need to be around to keep an eye on the big picture. So to know when to deploy automation or decommission it or redevelop a system as it adapts over time.

This means that there's going to be some new rules for supervision of these systems in newsrooms and in fact, these positions are already popping up at organizations that are using a lot of automated content.

So, at Reuters there's an automation bureau chief, at The Associated Press there's an automation editor, Bloomberg has hired several people to work on automation directly, and these types of roles will need editorial thinking as much as they'll need kind of almost a data science mindset or at least a capacity to understand data, and the capacity to understand the current state of the art of the technology.

And in addition to these types of high end positions, more editors and supervisors and overseers there may also be some growth in lower level or entry level positions. So the positions associated with maintaining templates or updating databases and data sets, these might be typically a little bit lower skill, and so we might also see a demand at that end of the market for people who can do some of the more custodial work related to keeping these systems in operation.

OK! So that's it for now. In the next video, I want to get into more detail on the benefits and limits of automated content, and then in the third video this week, I'll demo a template writing tool called Arria Studio so that you get a feel for how a system like this really works in practice. Thanks and I'll see you then.