Narrator

Generating Finished Intelligence Products From Structured Data
Who We Are

Sergey Polzunov
Software Engineer

Jörg Abraham
Chief Analyst
Agenda

- Setting the Scene
- From Structured Data to Narratives
- NLG in the Cyber Threat Intelligence Domain
- Narrator - A Proof-of-Concept
- Lessons Learned, Takeaways, Further Consideration
Setting The Scene
Natural Language Generation (NLG), which is a subfield of artificial intelligence and computational linguistics [...] can produce meaningful texts in English or other human languages from some underlying non-linguistic representation of information.

Widespread Use of NLG

- Real Time Traffic & Weather Forecast
- Journalism
- Business Intelligence
- Sports Reporting
- Property Bot
- Financial Analysis and Reporting
NLG in Journalism

The Washington Post experiments with automated storytelling to help power 2016 Rio Olympics coverage

The Washington Post, August 5, 2016, WashPostPR

Press Association wins Google grant to run news service written by computers

The Guardian, July 6 2017, Julia Gregory

EY, Deloitte And PwC Embrace Artificial Intelligence For Tax And Accounting

Forbes, November 14 2017, Adelyn Zhou
Cuomo ordains New Yorkers to wear face masks in public

In crowded neighborhoods, there will be a requirement for face coverings, while the New York City authorities said they would create emergency food reserves and take other measures to ensure supply for residents.

The face mask rule will take effect only three days later, providing people with enough caution to buy or find one.

The new mandate requires that a mask or mask should be fitted on busy roads, on public transport, or in situations where people cannot preserve their social separation.

De Blasio said the city does not just feed hungry New Yorkers but also ‘people to work.’

So far, the city already distributes 250.000 free meals a day to schools and provides elderly people 26.000 meals per day.

Monok.com, April 15 2020
NVIDIA Leverages NLG to Augment Marketing Analytics

Automated Insights’ Wordsmith platform integrated within Tableau
Databook - 5.8 M Executive Summaries Written Per Year

On page load, the executive summary is instantly written to reflect real-time data from the market.

Mastercard - Customer Habits & Buying Patterns

“This may be a key driver of your lower average purchase amount. You might consider additional promotional activity for lunch customers to come back for happy hour or dinner.”

From Structured Data to Narratives
From Structured Data to Narratives

Document Planner → Microplanner → Surface Realiser
Document Planner

- **Context determination** - Decide what information should be communicated out of the full dataset available
- **Document structuring** - How the information should be structured

→ Produces a *document plan* object
Lexicalization - Chose the words that will represent concepts

Referring expressions generation - chose proper names, pronouns and references

Aggregation - group information that should be expressed in one lexical block (phrase, paragraph, section)

→ Produces a text specification object
Surface Realiser

- Generating final text blocks from a text specification tree, produced by microplanner

→ Produces *final text*
NLG Systems & Services

- Association for Computational Linguistics (ACL)
  [https://aclweb.org/aclwiki/Downloadable_NLG_systems](https://aclweb.org/aclwiki/Downloadable_NLG_systems)

- Multiple commercial NLG services for various domains
NLG in the Cyber Threat Intelligence Domain
NLG in the Cyber Threat Intelligence Domain

- Automated email Generation for Targeted Attacks using Natural Language

Avisha Das, Rakesh Verma
Department of Computer Science University of Houston, Houston, Texas
NLG in the CTI Domain - The Challenge

Millions of structured information → Finished intel product is often a written report
How NLG Can Help Intelligence Analysts?

- Free up analyst’s time by automating what can be automated
- Expanded coverage
- Increases capability for investigations by looking at big data sets
- Consistency & conformance to standards

Augmentation rather than Replacement
Narrator - A Proof-of-Concept
Use STIX2 bundles as input data format.

Require high quality STIX2 with necessary relations / properties set.

Create draft report, which must then be further edited by an analyst.

Focus on producing routine factual sections of a document which human authors often find monotonous to write.
Proof-of-Concept Narrator - NLG for CTI

Overview of Donot Team.

The intrusion set’s name is “Donot Team” and it is also known as “Defect Team”, “SectorIoT2”, “APT-C-35” and “Intrusion Set: Donot Team”. The intrusion set’s description is:

Target: Government agencies in the Kashmir region.
The intrusion set has been first observed on August 16, 2018 and has been last seen on January 11, 2019. It is an organized group.

Motivation.

The set’s primary motivation is not-provided and its secondary motivation is political. The intrusion set’s goal is not-provided.

Associated campaigns.

It is related to 4 campaigns:

*Donot Team: Stealth Android malware campaign* campaign. It is also known as “Stealth”. The campaign’s objective is not-provided. Its description is:

Recently, researchers have observed a large-scale upgrade of its malicious Android APK framework to make it more stable and predictable. Since the new APK framework is different from the one used in the past, we named it as Stealth since “Stealth” is frequently used in the code.

First activity related to it is dates back to April 10, 2019 and the campaign’s last activity date back to July 12, 2019.

*Credentialed Harvesting Campaign Targeting South Asian Government and Defense Organizations* campaign. The campaign’s objective and description are not-provided. The campaign has been first seen on February 1, 2019 and it has been last seen on September 2, 2019.

*Donot 2018 campaign targeting Android devices* campaign. It is also known as “Donot 2018 campaign”. Its objective is not-provided. The campaign’s description is:

From July 2017 to at least August 2018, Donot Team has used a mobile RAT disguised as KNS Lite (Kashmir News Service) to target organizations in Pakistan.

First events associated with the campaign date back to August 16, 2018 and its last seen time is not provided.

*SectorIoT2 Spear phishing Campaign March-July 2019* campaign. It is also known as “SectorIoT2 2019 Campaign”. Its objective is not-provided. The campaign’s description is:

From March to July this year, the ThreatRecon team noticed a spear phishing campaign by the SectorIoT2 group going on against the Government of Pakistan and organizations there related to defense and intelligence. Spear phishing emails are sent to their victims via Excel XLS files, which also their victims to enable macros which will end up executing the downloader. Malicious document files were included in some documents purporting to be a registration for the Pakistan Air Force.

It has been first observed on February 28, 2019. The campaign’s last seen time is not provided.

Associated victims.

Pakistan Defense and Intelligence Organizations targeted victim with identity class "class". Its sectors are defense and government-national. Its description is:

On March to July this year, the ThreatRecon team noticed a spear phishing campaign by the SectorIoT2 group going on against the Government of Pakistan and organizations there related to defense and intelligence. Spear phishing emails are sent to their victims via Excel XLS files, which also their victims to enable macros which will end up executing the downloader. Malicious document files have been included in some documents purporting to be a registration for the Pakistan Air Force.

Recent (as of 2019) tactics, techniques, and procedures related to the campaign include:

- Targeted email campaigns
- Use of Excel files with macros
- Phishing emails with attachment
- Use of sophisticated social engineering
- Stealing sensitive information
Lessons Learned, Pitfalls and Takeaways
Lessons Learned

- Strict criteria on quality of the produced content required (style, structure, words used)
  - Limits the use of artificial neural networks (unpredictability)
  - Post-generation moderation is required

- STIX2 partially not fit for purpose, due to limited object properties

- Additional sanitization required (due to analysis / information was missing in structured data)
Takeaway

- NLG requires a matching use case.
- Start with realistic goals.
- Data must be structured enough.
- Requires significant SME and engineering investment.
- Leave neural networks for later, and use them for text style transfer, synonyms selection, data importance estimation, etc.
Further Consideration

- Algorithmic Transparency
- Algorithmic Bias
- Balancing Artificial and Human Intelligence
- Additional controls for sensitive information?
Thank You

Questions? - jorg@eclecticiq.com / sergey@eclecticiq.com