

# Implementation of Robot Journalism by Programming *Custombot* using Tokenization and Custom Tagging

Naeun Lee, and Kirak Kim, Taeseon Yoon

Hankuk Academy of Foreign Studies, Yongin, Gyeonggi, South Korea

janice.rhea@gmail.com, lecille@gmail.com, tsyoon@hafs.hs.kr

**Abstract**—The paper introduces a prototype of an algorithm that creates personalized news articles about IT and technology based on each personal preference for a specific theme, criteria, or element. When provided a specific personal preference, the algorithm *Custombot* analyses the data, derives the most appropriate topic that contains the most elements preferred by a person, and eventually produces a one and only news article on that topic. While processing and analysing data by inductive reasoning, *Custombot* considers the concepts of news angle and filter bubble. Text segmentation (tokenization) and custom tagging are two of the tasks that are used for the construction of this system, allowing to make custom tags and insert matching information in appropriate places. Its result of customized news article can serve as a new service provided by news organizations to satisfy each consumer's needs, and can also be a stepping stone in expanding the role or increasing the importance of robot journalism in a broad field of journalism.

**Keywords**— Custom Tagging, Filter Bubble, News Angle, Personalized News, Robot Journalism, Tokenization

## I. INTRODUCTION

Artificial intelligence (shortened as A.I.) is nowadays regarded as one of the major issues throughout the world. Correspondingly, the implementation of A.I. in various fields with purpose of producing advanced results can be commonly seen in recent society. Among the diverse fields that usually implement A.I. – ranging from programming and statistics to problem solving – the field that will be specifically focused on in this paper is robot journalism.

Exact definition for robot journalism is not yet provided, yet there are certainly some cases that concepts such as robot journalists or robot editors have been introduced in real news industry. One of the early concepts related to robot journalism was developed by Google, which introduced a “robot editor” in Google News Service. Stories were grouped based on a clustering algorithm that identifies related articles by analyzing many factors. Then, readers were sent directly to articles, bypassing the originated sites' home pages [1]. These “robot editors” even determined what should be the site's lead story, what should be displayed on the home page based on a “source credibility measure” [2].

More recent case of introducing “robot journalist” was in January, 2015, when the Associated Press declared that it had implemented the automated system for writing a story, and that the system has been publishing approximately 4,300 stories per quarter – which is 14 times more than the previous manual output of AP's reporters and editors [3].

Yet, until now, the range of the topic that the system can cover is quite limited. Natural disasters, economics, and sports are some of the few topics that robot journalists usually deal with. This paper, therefore, attempts to expand this range of the topics that robot journalists can cover by writing articles about IT and technology.

Idea of the system initiated from the phenomenon called “filter bubble.” Defined as a result of personalized search, filter bubble is the concept that was negatively viewed by some scholars such as Eli Pariser, as, they believed, it has a detrimental influence on individuals by narrowing their viewpoints. However, based on the idea that its negative effect is not that severe, we chose to focus on the opposite viewpoint against Pariser, believing that: when used properly, the system of personalized search and providing personalized service based on that will do more good than harm.

This system works with an algorithm named *Custombot* that can incessantly produce myriads of different news articles based on the same set of data, selecting only necessary information from the whole set depending on each personal preference. There certainly are some cases in media or news organizations of creating basic templates of news articles for respective news categories – in case of sports, such categories could be baseball, basketball, or football, etc. – and robot journalists writing stories about different events by only changing several variables in the pre-set template. However, until now, there has been no case (therefore no precedent) of robot journalists producing various different stories about the same topic and same sets of data, which is the very system that is introduced in this paper.

To sum up, as an effort to expand the range of the topic that robot journalists can deal with, and moreover as an effort to provide fully customized service for the readers, our paper introduces a prototype of an algorithm that produces personalized IT and technology news article for every user based on each personal preference for specific criteria.

Therefore, this paper will first analyze some related concepts which explain the purpose and basis of the algorithm. Then, the overall procedure or method of programming the algorithm will be introduced, which will be followed by the result of this procedure. For the result, two news articles each customized for two respective users are provided.

## II. RELATED WORKS & TERMS

There are three concepts that explain the motive and basis of the idea included in this system. News angle is usually regarded a major element that must be in consideration when writing a news article. Therefore, this “robot-journalistic” system must also have an ability to consider this element within its algorithm in order to produce stories that are in similar quality with the ones written by humans. Inductive reasoning is the main idea of the procedure that the algorithm undergoes while writing an article. Filter bubble, mentioned earlier, is the phenomenon that is generally viewed in negative perspective, which we object and fully put in use the positive aspect of the phenomenon in the system.

### A. News Angle

In the field of journalism, there is an element called “news angle,” which indicates of a news story which is highlighted and elaborated. It is the chosen perspective, emphasis, bias or focus from which a news item is told. News angles draw from diverse aspects of newsgathering circumstances, such as the medium through which an item is being relayed, or the news values by which it is drawn in as news. Objectivity demands that the angle be seen to arise from the apparent facts of the situation, rather than being consciously imposed by a journalist pursuing a specific agenda, through more partisan models of journalistic practice display news angles as proof of their perspective [4].

### B. Inductive Reasoning

Robot journalism can be regarded as an extension of computational journalism. According to Hamilton and Turner, computational journalism is “the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism” [5]. In other words, all activities of journalism that include computation can be considered as a part of computational journalism.

Computational journalism shows significant difference from the original journalism in its method of writing an article. Whereas deductive method of setting the conclusion first and then collecting the data that fits the conclusion is commonly used in journalism, inductive method is used in computational journalism: deriving conclusion after analyzing data.

This method used in computational journalism is the exact method that is used in the algorithm introduced in this paper. Based on the background data about diverse topics and the data of specific personal preference, the algorithm can derive a conclusion, determining the most appropriate topic that will interest a user the most. Hence, the product of this algorithm will be a personalized news article with considerable amount of information that fits the interest of a reader.

### C. Filter Bubble

Coined by internet activist Eli Pariser, term “filter bubble” indicates a result of a personalized search by a website algorithm. This website algorithm guesses what information a user would like to see based on the information or profile of each user, blocking the users from other new opinions other than their own ones. Pariser showed concern about this result that existence of filter bubble will eventually lead to isolation of individuals in their own cultural or ideological “bubble” with narrowed outlook [6].

Filter bubble is therefore known to be one of the possible dangers of robot journalism. Lee and Kim also were anxious about its aftermath; they claimed that that excessive usage of algorithm in journalism can lead people to be exposed only to information with limited perspectives and therefore become more biased [7].

However, even though the bubble effect shows negative implications for excessive personalized search, there were some precedents arguing that the effect itself is minimal and addressable. Zhang introduced several approaches that together composed the Auralist recommender, a system that explicitly balances the conflicting goals of accuracy, diversity, novelty, and serendipity. He mentioned that their techniques offer a simple and well-grounded way to diffuse the effects of so-called “filter bubble” [8].

For this case, filter bubble is definitely the main concept and main idea of the article, as the purpose of an algorithm is to sort out certain information that fits with the preference and write it in format of a news article. However, for this case, it could be anticipated that filter bubble will not give much harm or negative influence to readers. The very field of IT and technology does not have much variations in “outlooks” compared to other fields such as politics. This is due to the fact that whereas there are myriads of subjective ideas in fields such as politics that cannot be yielded into a single conclusion, the major ideas of IT or technology articles are based on objective data.

## III. NATURAL LANGUAGE PROCESSING

Natural Language Processing (shortened as NLP) is a field related to interactions between computers natural languages. Many tasks in NLP are the ones that can be applied in real life, if not merely subtasks that are used to aid in solving larger tasks. In this paper, natural language processing is the core technique for the formation of algorithm [9].

There are myriads of tasks in NLP, but there are some major tasks that are most commonly researched. One of the most prevalently known task is one that is called as Information Retrieval (IR), which is commonly used in search engines, as it is concerned with storing, searching, and retrieving information [10]. In similar sense, Information Extraction (IE) is concerned with extraction of semantic information from unstructured text and mere groups of data [11].

Other than extracting and retrieving information, some tasks deal with classifying the whole text, sentences, or components of sentences. Text Classification is the task which

computer classifies all the words according to the pre-set categories. Document Classification serves all the same functions, except that it classifies the category of the whole document [12]. Topic Recognition is another task which is in the similar term; it separates text segments and identify the topic of the text. Through Sentence Parsing, also called as Syntactic Analysis, recognition of sentence structures of the text can be automatically done [13].

There also exists some other tasks related to speech. Speech Recognition is the one that makes computer to understand human's natural language by determining the textual meaning of the speech. On the opposite, Text-to-Speech (TTS) is to convert text into speech in natural language. During this process, Speech Segmentation, which divide the whole audio clip into word segments, is fundamental [14]. Semantic Role Labelling is also used for Dialogue System; it is a task which identifies sentence elements such as subject and object in each sentence [15]. Furthermore, NLP even allows translation of language by using Machine Translation.

Among those, tasks mainly used for this paper are text segmentation and part-of-speech tagging.

#### A. Text Segmentation

Text segmentation, also called tokenization or word segmentation, is a process of dividing a string into several meaningful substrings, such as units of words, sentences, or themes [14]. In this paper, word and sentence segmentations are especially used to divide a string into its component words and sentences.

#### B. Part-of-speech Tagging

Part-of-speech tagging, also known as grammatical tagging or word-category disambiguation, is a process of identifying a word in a corpus as corresponding to a particular part of speech, based on its context and literal definition. It is a subtask of Text Segmentation.

#### C. Natural Language Toolkit (NLTK)

Natural Language Toolkit is a platform for Natural Language Processing in Python program [9] [16]. For this paper, this toolkit is used to produce sentence and word fragments, which is one of the significant parts of the overall algorithm.

### IV. METHOD & PROCEDURE

*Custombot* mainly undergoes six steps: text mining, text processing, tokenization, custom tagging, text analysis, and article writing. Additionally, before starting the whole process of analyzing and processing by using *Custombot*, personal preference should be first identified.

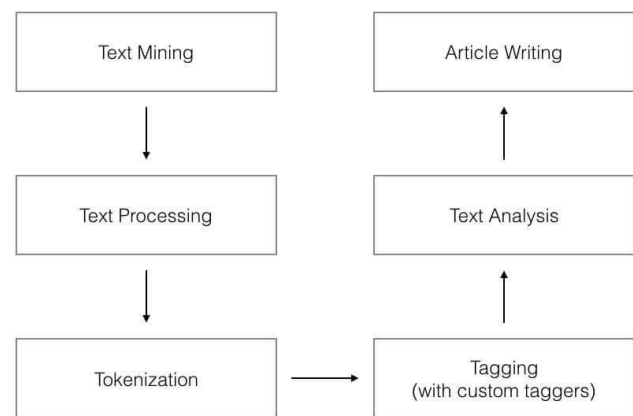


**Figure 1.** An algorithm *Custombot* which produces customized news articles for each individual based on the personal preference

#### A. Identifying Personal Preference

A user is initially required to identify his or her personal preference about a certain theme. This can be done by doing a short survey, asking criteria or elements that the individual deems as most important.

For this paper, we randomly set two users whose interests are in IT and technology. We hypothesized that both of them chose IT and technology as “interested theme,” and then we randomly set their personal preferences. User 1 is set to prefer cell phone that has a unique design, a solid body, a high-quality audio system, an expandable storage, and affordable price – but pay not much attention on Wi-Fi connectivity and processing speed. User 2 is set to consider features such as solidness of body, quality of camera lens, good audio system, removable battery, and expandable storage as essential, but view features such as size of storage, screen clearness, and speed comparatively trivial.

#### B. Text Mining

Two necessary elements for text mining are a text extractor and data source expressed in text. Every set of data that can be a potential content of the news article is extracted from a data source by using text extractor.

For this paper, a web API for the boilerpipe was used to extract text. It provides algorithms to detect and remove the surplus clutter around the main textual content of a web page, and eventually allows text mining [17].

In regard of the specific theme of IT and technology of the news article, a data source dealing with the very theme was necessary. Therefore, PCMag, a website in which information related to technology product is provided, was used as a data source. Profiles of smart phones produced by companies such as Apple, Asus, Google, Huawei, LG, Motorola, Samsung, Sony, and ZTE were extracted for this paper.

#### C. Text Processing using Regular Expression

Extracted text is read in Python program and is processed by using regular expression. Regular expression is a sequence of characters that define a search pattern, mainly for use in pattern matching with strings.

#### D. Tokenization using NLTK

With Natural Language Toolkit, extracted text is tokenized into units of words and sentences.

#### E. Tagging using Custom Tags

After creating some word and sentence fragments, the next step is to apply custom tags to those fragments. Tagging is a process of categorization of tokens based on custom tags.

Table 1 shows the example of custom tags that are used in order to produce stories. These custom tags are used for stories dealing with smart phones and related products. Explanations for every custom tag are provided on the table.

TABLE 1. LISTS OF SEVERAL CUSTOM TAGS AND THEIR EXPLANATIONS

|                  | Custom Tag    | Explanation                           |
|------------------|---------------|---------------------------------------|
| Hardware         | HWstrong      | Strong Hardware                       |
|                  | HWdesign      | Fancy Physical Design                 |
|                  | HWthin        | Slim and thin                         |
|                  | HWsmall       | Small size                            |
|                  | HWheavy       | Heavy and weighty                     |
| Audio            | AUgood        | High-quality audio system             |
|                  | AUbad         | Low-quality audio system              |
|                  | AUearphone    | Earphone jack exists                  |
|                  | AUearphonex   | No earphone jack                      |
| Battery          | BTremov       | Removable battery                     |
|                  | BTremovx      | Non-removable battery                 |
|                  | BTlong        | Long battery life                     |
|                  | BTshort       | Short battery life                    |
| Camera           | CMdual        | Dual camera                           |
|                  | CMhqlens      | High-quality camera lens              |
|                  | CMlqlens      | Low-quality camera lens               |
|                  | CMlowlight    | Camera for Low-light photography      |
| User Convenience | UCwaterprf    | Waterproof                            |
|                  | UCfingerprntx | No fingerprint verification sensor    |
|                  | UCwaterprfx   | Not watertight                        |
|                  | UCwlchargex   | Wireless charging not possible        |
|                  | UCfingerprint | Fingerprint recognition sensor exists |
| Screen           | SCclear       | Clear screen                          |
|                  | SC720p        | 720p screen                           |
|                  | SC1080p       | 1080p screen                          |
|                  | SCbright      | Bright screen                         |
|                  | SCdim         | Dim screen.                           |
|                  | SCwide        | Wide screen                           |
|                  | SCshatterprf  | Shatterproof.                         |
|                  | SChighres     | High screen resolution                |
|                  | SClowres      | Low screen resolution                 |
| Price            | PRlow         | Cheap                                 |
|                  | PRhigh        | Expensive                             |
| Storage          | STGexpand     | Expandable storage                    |
|                  | STGexpandx    | Not expandable storage                |
|                  | STGsmall      | Small storage                         |

|             |              |                           |
|-------------|--------------|---------------------------|
| Performance | MEbig        | Large memory capacity     |
|             | COgood       | Fair connectivity         |
|             | WFbad        | Unstable Wi-Fi connection |
|             | SWbloatware  | Bloatware exists          |
|             | SWbloatwarex | No bloatware              |
|             | SWsupportx   | Dated software            |
|             | NWdb         | Dual band network         |
|             | NWdbx        | non-dual band network     |
|             | PFbad        | Low performance           |
|             | SPhigh       | Strong processing power   |
|             | SPlow        | Weak processing power     |

#### F. Text Analysis

Among these tokens, there will be certain tokens which match with the personal preference of a certain user. Hence, after applying custom taggers, the numbers of tokens that fit a certain personal preference should be identified and saved for each text document. The algorithm then selects the text with the highest token number and uses it as basic information to produce an article.

#### G. Producing Articles

Based on the pre-set frame of the article, the algorithm produces articles by putting appropriate tokens on specific places indicated with respective custom taggers. In order to diversify the organization and structure of the news articles, a lot of word and sentence tokens which are considered as synonyms or which have similar meanings should be assigned for a single custom tagger.

For this paper, average of four tokens were assigned to one custom tagger, allowing the algorithm to randomly apply one of those tokens in the places where that certain custom tagger is identified.

### V. RESULT

Product of the algorithm is the news article that listed several sentences with information about each custom tags. Two random users with unique personal preferences can get different news articles from this algorithm.

The first customized news article is made for the random user User 1. User 1 prefers smartphone with a unique design, a solid body, a high-quality audio system, an expandable storage, and affordable price – but pays not much attention on its Wi-Fi connectivity and processing speed. Based on this preference, the algorithm analyzed data of every product on smartphone market and chose the most optimized product for User1: ZTE Axon 7. News article presented below is the story only for User 1.

*“ZTE-Axon-7 is on the market this week. It costs \$399.99. This product is well-known for its sturdiness. This product has a high-quality audio system. It is comparatively low in price. It has fingerprint verification sensor. Overall design of the product is regarded as unique. This product is able to expand its storage. Its network is dual-band. It has a low-light camera. Weight is one of the shortages of ZTE-Axon-7; it is quite weighty. Quality of camera lens is not very good.”*

The second customized news article is made for another random user named User 2. User 2 regards some features such as hardware, quality of camera lens, good audio system, removable battery, and expandable storage as essential, but view features such as size of storage, screen clearness, and speed as comparatively trivial. The most optimized product that the algorithm found out was LG V20. Below is the personalized article produced for User 2.

*"LG\_V20 is on the market this week. It costs \$672.00. It has a sturdy body. It guarantees a good call quality; its quality of transmission, earpiece volume, and background noise cancellation is fair enough. LG\_V20 has a removable battery. Functions of camera are diverse, which is one of the advantages of LG\_V20. LG\_V20 is able to expand storage. This product is not watertight. Wireless charging is not available for this product. There are a lot of unnecessary features in the software. Weight is one of the shortages of LG\_V20; it is quite weighty."*

When juxtaposing these two articles, it is conspicuous that order of sentences, structure of sentences, and usage of words are all different from one another.

## VI. CONCLUSIONS & FUTURE WORKS

*Custombot* allows an individual to get customized news article when the user's preference is chosen and input to the algorithm. Depending on the categories that a user choose, the content, structure, and order will be differed for every article.

Our next goal would be to further develop the algorithm to write articles indistinguishable from the articles written by humans by (1) adding an algorithm of machine learning in the process in order to analyze more data and diversify forms of sentences, eventually improving the quality of sentences tantamount to the quality of ones written by human journalists and (2) using sentimental analysis, one of the major tasks of natural language learning, for other algorithms which will write articles about other topics and themes such as social events, in which usage of vocabularies containing emotions are inevitable.

Our ultimate goal in the future is to achieve absolute automatization by finding a method to acknowledge each user's preference by using additional algorithms such as machine learning and back propagation algorithm. Further support of our model will consolidate the foundation of the development of robot journalism and expansion of the role of artificial intelligence. Customized news articles by *Custombot* will offer a new form of article in field of journalism.

## REFERENCES

- [1] (2002) L. Walker, "Google News, Untouched by Human Hands," *The Washington Post*, [Online]. Available: <http://www.washingtonpost.com>
- [2] (2002) H. Kurtz, "Robotic Journalism: Google Introduces Human-less News," *The Washington Post*.
- [3] (2015) F. Gleyo, "AP Has A Robot Journalist That Writes A Thousand Articles Per Month," *Tech Times*, [Online]. Available: <http://www.techtimes.com>
- [4] B. Zelizer and S. Allan, *Keywords In News And Journalism Studies*, Maidenhead, Berkshire, England: Open UP, McGraw Hill Education, 2010, p. 6.
- [5] J. T. Hamilton and F. Turner, "Accountability through algorithm: Developing the field of computational journalism," *Communications of the ACM*, 2009.
- [6] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, New York: Penguin, 2011.
- [7] D. Kim and J. Lee, "Robot Journalism: Algorithmic Approach to Automated News Article Generation," *Korean Society For Journalism And Communication Studies*, 2015.
- [8] Y. C. Zhang, D. Séaghdha, D. Quercia, and T. Jambor, "Auralist: Introducing Serendipity into Music Recommendation," *WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 13-22, 2012.
- [9] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, Beijing: O'Reilly, 2009.
- [10] S. Dominich, *The Modern Algebra of Information Retrieval*, Berlin: Springer, 2008.
- [11] U. M. Munshi and B. B. Chaudhuri, *Multimedia Information Extraction and Digital Heritage Preservation*, Singapore: World Scientific, 2011.
- [12] D. Sarkar, *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data*, Berkeley, CA: Apress, 2016.
- [13] A. Sanfeliu, J. F. Martinez Trinidad, and Jesús A. Carrasco Ochoa, *Progress in Pattern Recognition, Image Analysis and Applications: 9th Iberoamerican Congress on Pattern Recognition, CIARP 2004, Puebla, Mexico, October 26-29, 2004: Proceedings*, Berlin: Springer, 2004.
- [14] P. Taylor, *Text-to-Speech Synthesis*, Cambridge, U. K.: Cambridge University Press, 2009.
- [15] J. L. Vicedo, *Advances in Natural Language Processing: 4th International Conference, EsTAL*, Berlin: Springer, 2004.
- [16] (2015) The NLTK website. [Online]. Available: <http://www.nltk.org/>
- [17] Web API for the boilerpipe Java library. [Online] Available: <https://boilerpipe-web.appspot.com/>



**Naeun Lee** was born in Suwon, South Korea, in 1999. This author is currently studying at Hankuk Academy of Foreign Studies in Yongin, Gyeonggi. This author has worked as a student science journalist for National Science Museum in Korea in 2016. From 2017, she will work as an IT/science journalist for Korea Youth Press Corps. Highly interested in computer science and robot journalism, she wishes to study artificial intelligence, bioinformatics, or science journalism in the future.



**Kirak Kim** was born in Cheongju, South Korea in 1999. This author is currently studying at Hankuk Academy of Foreign Studies in Yongin, Gyeonggi. He wishes to study computer science and engineering in the future.



Taeseon Yoon was born in Seoul, Korea, in 1972. He was Ph.D. Candidate degree in Computer education from the Korea University, Seoul, Korea, in 2003. From 1998 to 2003, he was with EJB analyst and SCJP. From 2003 to 2004, he joined the Department of Computer Education, University of Korea, as a Lecturer and Ansan University, as an Adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies in Yongin, Gyeonggi, where he is a Computer Science and Statistics Teacher. He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-do, Korea, 2013.