Article Summarization Using Natural Language Processing

Shubham Rai

Department of Computer Science, The University of Texas at Dallas

Richardson, TX 75080 USA

Email: scr130130@utdallas.edu

**Abstract**

This paper introduces an approach to summarize any article, text based on subject–verb–object (SVO) in a sentence structure, where the subject comes first, the verb second, and the object third.  The Article summarizer application will take any text: news article, blog posts, reviews etc. entered by the user and find the Subject (i.e. actors, present in the text), Verbs (action performed by the actor, object) and Object Structure (keywords surrounding the objects). SVO is the most common order observed in terms of number of speakers.  A simple sentence is built from one independent clause. A compound and complex sentence is built from at least two clauses. Also, clause must have subject and verb. So our task is to split an articles into different sentence and then into clauses. Once we get the pair of subject and verb, then get all dependencies that have link to the subject, then extract unique word from these dependencies. Using this technique, the result declare the Subject-Verb-Object structure present in each paragraph of the article and the surrounding words present around the actors, verbs; forming a summary of each paragraph.

**1. Introduction**

Summarization is the process of reducing a text document in order to create a summary that retains the most important points of the original document. Generally, there are two approaches to automatic summarization: extraction and abstraction. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate.

While extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extraction of sentences, abstractive summarization puts strong emphasis on the form, aiming to produce a grammatical summary, which usually requires advanced language generation techniques.

The intention of the application is to summarize an article (document), where the goal is to select whole sentences to create a short summary of the paragraph. To identify an actor, verb and object we make use of Stanford NER which is a java implementation of a Named Entity Recognizer. Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names. An actor can be a person, organization, a location or words present in lexicon (dictionary). In this application we use a 7 class model : Time, Location, Organization, Person, Money, Percent, Date. The  Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

The article is divided into various paragraphs, the application will then go through the text line by line and allocate the part of speech tags to each word. The NER will help in determining the actors which can be a person, organization or a word in the lexicon. It will also determine the verbs and the surrounding words. This will help us in forming a Subject Verb Object structure for each paragraph. The summary will also provide several keywords such as words present in the title that were also in the paragraph, date, location, money, cardinal numbers etc..

**2. Prior research**

Most early work on single-document summarization focused on technical documents. Perhaps the most cited paper on summarization is that of (Luhn, 1958), that describes research done at IBM in the 1950s. In his work, he proposed that the frequency of a particular word in an article provides a useful measure of its significance. As a first step, words were stemmed to their root forms, and stop words were deleted. Luhn then compiled a list of content words sorted by decreasing frequency, the index providing a significance measure of the word. On a sentence level, a significance factor was derived that reflects the number of occurrences of significant words within a sentence, and the linear distance between them due to the intervention of non-significant words. All sentences are ranked in order of their significance factor, and the top ranking sentences are finally selected to form the auto-abstract.

Edmundson (1969) described a system that produces document extracts. His primary contribution was the development of a typical structure for an extractive summarization experiment. Two features were used: the presence of cue words (presence of words like significant, or hardly), and the skeleton of the document (whether the sentence is a title or heading). Weights were attached to each of these features manually to score each sentence. During evaluation, it was found that about 44% of the auto-extracts matched the manual extracts

The document summarization technology has been widely studied in natural language processing and has led to several real time advanced web-based systems that are currently available some of which have been listed below.

Ultimate Research Assistant- performs text mining on Internet search results to help summarize and organize them and make it easier for the user to perform online research. Specific text mining techniques used by the tool include concept extraction, text summarization, hierarchical concept clustering (e.g., automated taxonomy generation), and various visualization techniques, including tag clouds and mind maps.

iResearch Reporter- Commercial Text Extraction and Text Summarization system, free demo site accepts user-entered query, passes it on to Google search engine, retrieves multiple relevant documents, produces categorized, easily readable natural language summary reports covering multiple documents in retrieved set, all extracts linked to original documents on the Web, post-processing, entity extraction, event and relationship extraction, text extraction, extract clustering, linguistic analysis, multi-document, full text, natural language processing, categorization rules, clustering, linguistic analysis, text summary construction tool set.

Newsblaster-is a system that helps users find news that is of the most interest to them. The system automatically collects, clusters, categorizes, and summarizes news from several sites on the web (CNN, Reuters, Fox News, etc.) on a daily basis, and it provides users an interface to browse the results.

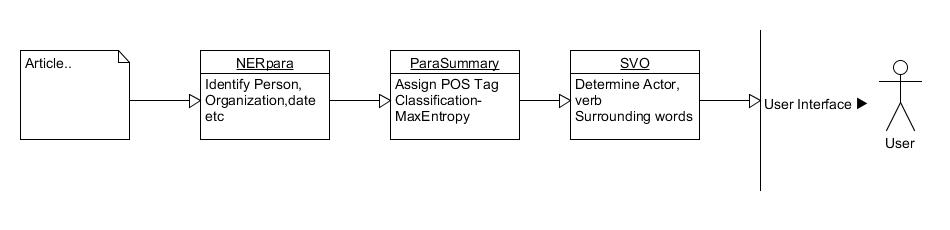
NewsInEssence- may be used to retrieve and summarize a cluster of articles from the web. It can start from a URL and retrieve documents that are similar, or it can retrieve documents that match a given set of keywords. NewsInEssence also downloads news articles daily and produces news clusters from them.

NewsFeed Researcher- is a news portal performing continuous automatic summarization of documents initially clustered by the news aggregators (e.g., Google News). NewsFeed Researcher is backed by a free online engine covering major events related to business, technology, U.S. and international news. This tool is also available in on-demand mode allowing a user to build a summaries on selected topics.

Scrape- This is like a search engine, but instead of providing links to the most relevant websites based on a query, it scrapes the pertinent information off of the relevant websites and provides the user with a consolidated multi-document summary, along with dictionary definitions, images, and videos.

Inall of the above projects/applications, articles on the same story from various sources are presented together and summarized using state-of-the-art techniques. This application summarizes an article based on the SVO structure and also presents to the user various key details like organization, money, people, date, cardinal numbers, time etc. mentioned in each paragraph.

**3. System Design**



The application consist of four main classes ParaSummary, NERpara, ProcessSummary and Svo.

The ParaSummary manages all the methods for each paragraph, it also handles the POS Tagger of Stanford. It uses the class NERpara to handle the NER Library of Stanford. For example it return the verbs of the paragraph, or the words on the lexicon. The paraSummary is the most important part of the application as it extracts all the key details i.e. actors, verbs and object needed to formulate the subject-verb-object structure. It makes use of Maximum Entropy model and simultaneously uses many contextual "features" to predict the POS tag.

The NERpara class uses the Stanford Named Entity Recognizer. Named Entity Recognition (NER) labels sequences of words in a text which recognizes named (person, location, organization, misc.), numerical (money, number, ordinal, percent), and temporal (date, time, duration, set) entities.

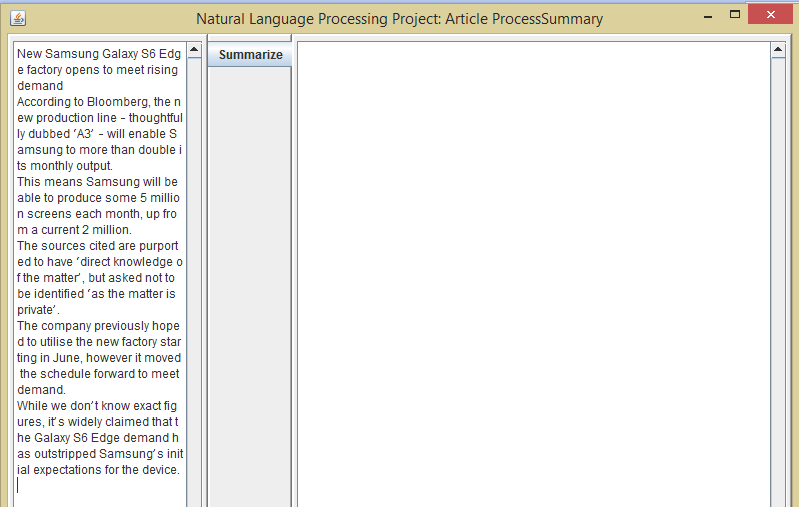
The ProcessSummary class uses the classifier which is a machine learning tool that will take data items and place them into one of *k* classes. A probabilistic classifier, like the one used, can also give a probability distribution over the class assignment for a data item. This software is a implementation of a maximum entropy classifier. Maximum entropy models are otherwise known as conditional loglinear models, and are essentially equivalent to multiclass logistic regression models (though parameterized slightly differently, in a way that is advantageous with sparse explanatory feature vectors). This class also checks the keywords extracted in the NERpara and ParaSummary against a lexicon list.

The Svo class finds Subject, Verb, Object Structure and words surrounding the Object. It uses the information provided by ProcessSummary and NERpara to determine the actors and forms a summary around the actor by choosing the verbs and surrounding words directly related to the actor. If there are multiple actors, then it will check for words surrounding it and also if the actor is present in the title. A summary will be formed by taking in the main subject, verb and object.

The rest of the classes Main, btnSummary, Article and output Summary deal with the gui part of the application.

**4. Input/Output**

On the left hand side we input the article and on right we get the desired output

Input:

New Samsung Galaxy S6 Edge factory opens to meet rising demand

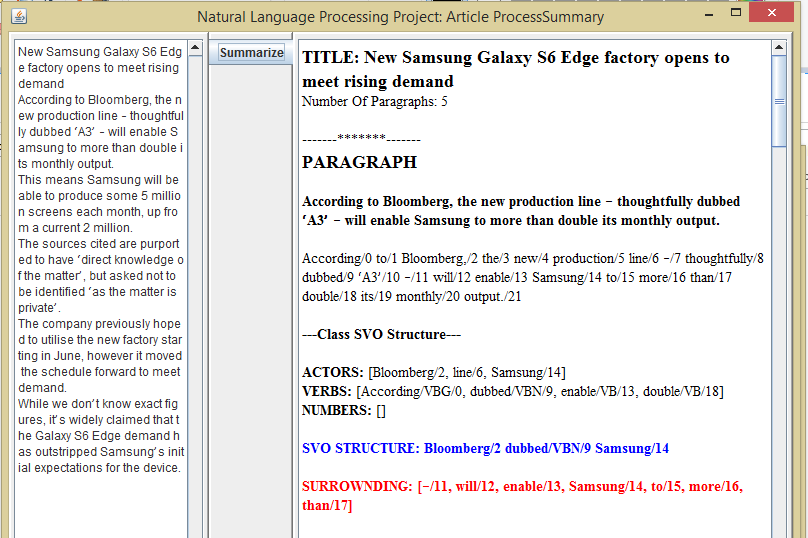
According to Bloomberg, the new production line – thoughtfully dubbed ‘A3’ – will enable Samsung to more than double its monthly output.

This means Samsung will be able to produce some 5 million screens each month, up from a current 2 million.

The sources cited are purported to have ‘direct knowledge of the matter’, but asked not to be identified ‘as the matter is private’.

The company previously hoped to utilise the new factory starting in June, however it moved the schedule forward to meet demand.

While we don’t know exact figures, it’s widely claimed that the Galaxy S6 Edge demand has outstripped Samsung’s initial expectations for the device.

Output

Full Output:

**TITLE: New Samsung Galaxy S6 Edge factory opens to meet rising demand**  
Number Of Paragraphs: 5  
  
-------\*\*\*\*\*\*\*-------  
**PARAGRAPH**  
  
**According to Bloomberg, the new production line – thoughtfully dubbed ‘A3’ – will enable Samsung to more than double its monthly output.**  
  
According/0 to/1 Bloomberg,/2 the/3 new/4 production/5 line/6 –/7 thoughtfully/8 dubbed/9 ‘A3’/10 –/11 will/12 enable/13 Samsung/14 to/15 more/16 than/17 double/18 its/19 monthly/20 output./21  
  
**---Class SVO Structure---**  
  
**ACTORS:** [Bloomberg/2, line/6, Samsung/14]  
**VERBS:** [According/VBG/0, dubbed/VBN/9, enable/VB/13, double/VB/18]  
**NUMBERS:** []  
  
**SVO STRUCTURE: Bloomberg/2 dubbed/VBN/9 Samsung/14**  
  
**SURROWNDING: [–/11, will/12, enable/13, Samsung/14, to/15, more/16, than/17]**  
  
---Keywords---  
  
**Words on Title:** [new, samsung, to, to]  
**Words on Lexicon:** [line/6]  
**Verbs:** [According/VBG/0, dubbed/VBN/9, enable/VB/13, double/VB/18]  
**Organizations:** [Bloomberg/2, Samsung/14]  
**Cardinal Numbers:** []  
**People:** []  
**Money:** []  
**Percentages:** []  
**Dates:** []  
**Locations:** []  
**Time:** []  
**Adjectives:** [new/JJ/4, monthly/JJ/20]  
  
  
-------\*\*\*\*\*\*\*-------  
**PARAGRAPH**  
  
**This means Samsung will be able to produce some 5 million screens each month, up from a current 2 million.**  
  
This/0 means/1 Samsung/2 will/3 be/4 able/5 to/6 produce/7 some/8 5/9 million/10 screens/11 each/12 month,/13 up/14 from/15 a/16 current/17 2/18 million./19  
  
**---Class SVO Structure---**  
  
**ACTORS:** [Samsung/2]  
**VERBS:** [means/VBZ/1, be/VB/4, produce/VB/7]  
**NUMBERS:** [2, 5, 2 million/10, 2 million/18]  
  
**SVO STRUCTURE: Samsung/2 produce/VB/7 null**  
  
**SURROWNDING: [be/4, able/5, to/6, produce/7, some/8, 5/9, million/10]**  
  
---Keywords---  
  
**Words on Title:** [samsung, to]  
**Words on Lexicon:** []  
**Verbs:** [means/VBZ/1, be/VB/4, produce/VB/7]  
**Organizations:** [Samsung/2]  
**Cardinal Numbers:** [5, million, 2, million]  
**People:** []  
**Money:** [2 million/10, 2 million/18]  
**Percentages:** []  
**Dates:** []  
**Locations:** []  
**Time:** []  
**Adjectives:** [able/JJ/5, current/JJ/17]  
  
  
-------\*\*\*\*\*\*\*-------  
**PARAGRAPH**  
  
**The sources cited are purported to have ‘direct knowledge of the matter’, but asked not to be identified ‘as the matter is private’.**  
  
The/0 sources/1 cited/2 are/3 purported/4 to/5 have/6 ‘direct/7 knowledge/8 of/9 the/10 matter’,/11 but/12 asked/13 not/14 to/15 be/16 identified/17 ‘as/18 the/19 matter/20 is/21 private’./22  
  
**---Class SVO Structure---**  
  
**ACTORS:** [asked/13]  
**VERBS:** [cited/VBD/2, are/VBP/3, have/VB/6, asked/VBD/13, be/VB/16, identified/VBN/17, is/VBZ/21]  
**NUMBERS:** []  
  
**SVO STRUCTURE: asked/13 identified/VBN/17 null**  
  
**SURROWNDING: [not/14, to/15, be/16, identified/17, ‘as/18, the/19, matter/20]**  
  
---Keywords---  
  
**Words on Title:** [to, to]  
**Words on Lexicon:** [asked/13]  
**Verbs:** [cited/VBD/2, are/VBP/3, have/VB/6, asked/VBD/13, be/VB/16, identified/VBN/17, is/VBZ/21]  
**Organizations:** []  
**Cardinal Numbers:** []  
**People:** []  
**Money:** []  
**Percentages:** []  
**Dates:** []  
**Locations:** []  
**Time:** []  
**Adjectives:** [purported/JJ/4]  
  
  
-------\*\*\*\*\*\*\*-------  
**PARAGRAPH**  
  
**The company previously hoped to utilise the new factory starting in June, however it moved the schedule forward to meet demand.**  
  
The/0 company/1 previously/2 hoped/3 to/4 utilise/5 the/6 new/7 factory/8 starting/9 in/10 June,/11 however/12 it/13 moved/14 the/15 schedule/16 forward/17 to/18 meet/19 demand./20  
  
**---Class SVO Structure---**  
  
**ACTORS:** [company/1, forward/17]  
**VERBS:** [hoped/VBD/3, utilise/VB/5, starting/VBG/9, moved/VBD/14, meet/VB/19]  
**NUMBERS:** []  
  
**SVO STRUCTURE: company/1 hoped/VBD/3 forward/17**  
  
**SURROWNDING: [moved/14, the/15, schedule/16, forward/17, to/18, meet/19, demand./20]**  
  
---Keywords---  
  
**Words on Title:** [new, factory, to, to, meet, demand]  
**Words on Lexicon:** [company/1, forward/17]  
**Verbs:** [hoped/VBD/3, utilise/VB/5, starting/VBG/9, moved/VBD/14, meet/VB/19]  
**Organizations:** []  
**Cardinal Numbers:** []  
**People:** []  
**Money:** []  
**Percentages:** []  
**Dates:** [June/11]  
**Locations:** []  
**Time:** []  
**Adjectives:** [new/JJ/7]  
  
  
-------\*\*\*\*\*\*\*-------  
**PARAGRAPH**  
  
**While we don’t know exact figures, it’s widely claimed that the Galaxy S6 Edge demand has outstripped Samsung’s initial expectations for the device.**  
  
While/0 we/1 don’t/2 know/3 exact/4 figures,/5 it’s/6 widely/7 claimed/8 that/9 the/10 Galaxy/11 S6/12 Edge/13 demand/14 has/15 outstripped/16 Samsung’s/17 initial/18 expectations/19 for/20 the/21 device./22  
  
**---Class SVO Structure---**  
  
**ACTORS:** [claimed/8]  
**VERBS:** [know/VB/3, claimed/VBN/8, has/VBZ/15, outstripped/VBN/16]  
**NUMBERS:** []  
  
**SVO STRUCTURE: claimed/8 has/VBZ/15 null**  
  
**SURROWNDING: [S6/12, Edge/13, demand/14, has/15, outstripped/16, Samsung’s/17, initial/18]**  
  
---Keywords---  
  
**Words on Title:** [galaxy, s6, edge, demand]  
**Words on Lexicon:** [claimed/8]  
**Verbs:** [know/VB/3, claimed/VBN/8, has/VBZ/15, outstripped/VBN/16]  
**Organizations:** []  
**Cardinal Numbers:** []  
**People:** []  
**Money:** []  
**Percentages:** []  
**Dates:** []  
**Locations:** []  
**Time:** []  
**Adjectives:** [exact/JJ/4, initial/JJ/18]

**5. Conclusion**

The Application summarizes each paragraph of the article based on the Subject-Verb-Object structure.

The application is accurate in determining the actors present in the paragraph of the article

The Summary produced by the application is not entirely accurate, but it provides the basic essence of each paragraph of the article and the keywords present in it.

**6. References**

1. http://www-nlpir.nist.gov/projects/duc/index.html
2. http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf
3. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159–165. [2, 3, 6, 8]
4. Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2):264–285. [2, 3, 4]
5. http://nlp.stanford.edu/software/index.shtml
6. http://en.wikipedia.org/wiki/Automatic\_summarization
7. http://en.wikipedia.org/wiki/Multi-document\_summarization