WiseReporter: A Korean Report Generation System

Yunseok Noh and **Su Jeong Choi** and **Seong-Bae Park** and **Se-Young Park**
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
{ysnoh,sjchoi,sbpark}@sejong.knu.ac.kr seyoung@knu.ac.kr

Abstract

We demonstrate a report generation system called *WiseReporter*. The WiseReporter generates a text report of a specific topic which is usually given as a keyword by verbalizing knowledge base facts involving the topic. This demonstration does not demonstate only the report itself, but also the processes how the sentences for the report are generated. We are planning to enhance WiseReporter in the future by adding data analysis based on deep learning architecture and text summarization.

1 Introduction

The necessity of well-organized information about emerging topics grows fast, but the conventional search engines such as Google or Bing provide just a list of relevant documents. Since the results of the search engines are unstructured, there should be additional and expensive cost to provide users with exact information. However, due to extremely large volume of information amount, it is nearly impossible for users themselves to look over all contents and get the insight of topics of interest from them. From this point of view, we argue the need of a tool which enables analyzing a large volume of documents and summarizing them as a report that can be easily understood by the users.

As the very first step of the report generation tool, we demonstrate a prototype system called *WiseReporter* that translates knowledges in a knowledge base (KB) to text reports. There exist many large scale KBs such as Freebase and DBpedia, and several algorithms to add knowledges from web documents into a KB automatially (Carlson et al., 2010). Therefore, in this demonstration of WiseReporter, we focus only on the verbalization of the facts in a KB to generate a report for a specific topic.

Basically, WiseReporter is a template based generation system (Mellish et al., 2006). This approach has been broadly used for generating natural language texts from KB facts (Nadjet et al., 2014), where a KB fact consists of a relation and two entities linked by the relation. If there are natural language templates for every relation in a KB, then the facts can be easily transformed into natural language sentences by filling slots of the proper template with the entities of the facts. In addition to the templates and template-slot-filling, WiseReporter contains many processes for report generation such as macro- and micro-planning, and surface-form realization.

In this demonstration, we use two KBs for text report generation. One is a manually-constructed domain-specific KB associated with IT products, and the other is DBpedia to cover more general topics. With the KBs, we prove that WiseReporter provides reasonable results in terms of *linguistic quality* evaluation of DUC task, and also demonstrate how sentences of a report are generated from the KBs by visualizing some generation rules.

2 Overview of WiseReporter

2.1 System Architecture

WiseReporter adopts a pipelined architecture for natural language generation following several studies on ontology verbalization (Androutsopoulos et al., 2013; Nadjet et al., 2014). The architecture is typically composed with three major modules (Mellish et al., 2006): (i) text planning (also referred to as macro-planning), (ii) sentence planning (also known as micro-planning), and (iii) surface-form realization. The text planning module is responsible for choosing *what to say* and organizing the selected content in a coherent way. The sentence planning module is re-

The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations, pages 9–12,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

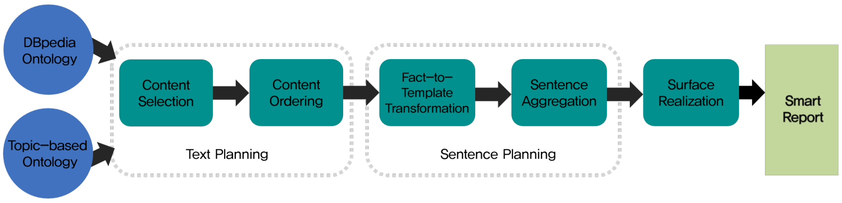


Figure 1: The overall pipelined architecture of WiseReporter.

sponsible for mapping the text plan to a linguistic structure, grouping information into sentences, and performing aggregation and lexicalization. At last, the surface-form realizer is in charge of rendering each sentence plan into a sentence string.

WiseReporter actually implements the pipelined architecture with the following five components (see Figure 1).

- **Content selection**. Both open planning and closed planning (Nadjet et al., 2014) are used for IT product KB and DBpedia respectively.
- **Content ordering**. Following the work of Androutsopoulos et al. (2013), we focus on enhancing local coherence by considering smooth topic change among adjacent sentences.
- **Fact-to-template transformation**. The rules for fact-to-template transformation are carefully designed with the consideration for Korean language phenomena such as Subject-Object-Verb (SOV) word-order and decision of postposition (*josa*) (Yang, 1995).
- **Sentence aggregation**. Several aggregation rules are applied to merging multiple simple sentences into a complex sentence. This step allows various and fluent natural language expressions.
- **Surface realization**. This component deals with several issues about realization of Korean surface-form including the problem of determining verbal endings (Yang, 1995). Especially, the endings related to tense, conjunction, and passive expressions are treated intensively.

The generated report consists of a number of paragraphs and an image related to the report topic. The image is inserted for helping users understand the generated texts better, and this image is simply obtained by Google image search. The typical layout of the report is pre-defined in this version of WiseReporter. The automatic layout arrangement and the appropriate image selection (or generation) remain as our future work.

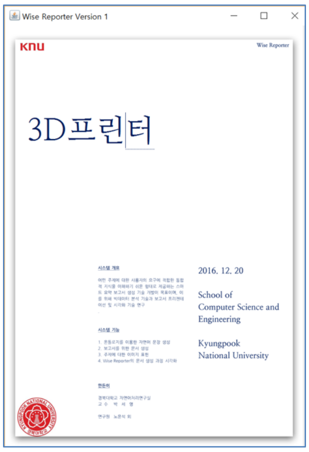
We evaluated the quality of the generated reports by human judgement. The linguistic quality evaluation for summarization was taken from the previous study of Over et al. (2007). Five native evaluators were asked to score the generated reports from 1 to 5 points on five evaluation items of grammaticality, non-redundancy, structure and coherence, referential clarity, and focus (Over et al., 2007). The grand average score on 10 reports was 3.6 of 5.0. This result is competitive with the average score of 1.96 at the work of Androutsopoulos et al. (2013)[1].

2.2 Knowledge Bases

WiseReporter makes use of two KBs for report generation. One is a domain-specific ontology constructed manually, and the other is DBpedia. Domain-specific ontologies are usually designed to represent specifications of a subject, but this is not enough for report generation to deliver information such as background or related events. Therefore, an ontology is designed that describes IT products in detail by analyzing the documents on IT products. For this, a number of natural language patterns are collected, and then many facts are harvested from the documents by pattern matching the patterns with the documents (Gerber and Ngomo, 2012). After that, the collected facts are refined manually for accuracy. The final IT product KB contains 239 facts in total.

In addition to domain-specific ontology in hand, DBpedia is also included in WiseReporter for wide coverage of the system. In order to generate reports for the topics in DBpedia, we defined 167 templates. After all, WiseReporter can produce reports on 29,255 different topics.

[1] This study evaluted their system by 1 to 3 scale on English texts from Wine Ontology.



Wise Reporter Version 1

KNU

Wise Reporter

3D프린터

시스템 개요

어떤 주제에 대한 사용자의 요구에 적합한 통합적 지식을 이해하기 쉬운 형태로 제공하는 스마트 요약 보고서 생성 기술 개발이 목표이며, 이를 위해 빅데이터 분석 기술과 보고서 프리젠테이션 및 시각화 기술 연구

시스템 기능

1. 온톨로지를 이용한 자연어 문장 생성
2. 보고서를 위한 문서 생성
3. 주제에 대한 이미지 표현
4. Wise Reporter의 문서 생성 과정 시각화

만든이

경북대학교 자연어처리연구실
교 수 박 세 영

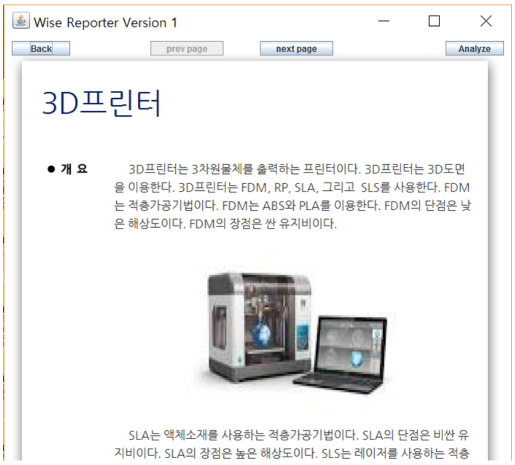
연구원 노은식 외

2016. 12. 20

School of
Computer Science and
Engineering

Kyungpook
National University

Figure 2: The report cover interface with the title ‘3D 프린터 (3D printer)’ inputed by a user.



Wise Reporter Version 1

3D프린터

● 개 요

3D프린터는 3차원물체를 출력하는 프린터이다. 3D프린터는 3D도면을 이용한다. 3D프린터는 FDM, RP, SLA, 그리고 SLS를 사용한다. FDM는 적층가공기법이다. FDM는 ABS와 PLA를 이용한다. FDM의 단점은 낮은 해상도이다. FDM의 장점은 싼 유지비이다.

SLA는 액체소재를 사용하는 적층가공기법이다. SLA의 단점은 비싼 유지비이다. SLA의 장점은 높은 해상도이다. SLS는 레이저를 사용하는 적층

Figure 3: The report about ‘3D printer’ generated by WiseReporter. This result is generated by using our IT product KB.

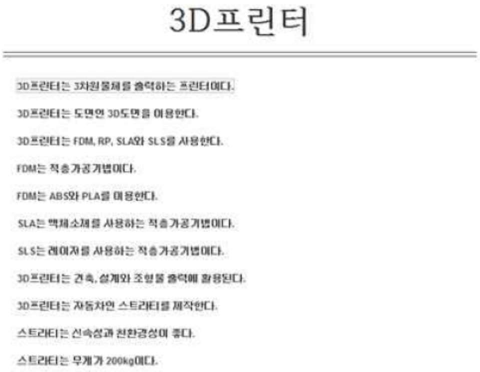
3 Outline of Demonstration

The following three steps outline our demonstration:

1. WiseReporter accepts a keyword such as ‘3D 프린터 (3D printer)’ in the title position of the report cover interface (see Figure 2).

2. WiseReporter returns a text report including an image about the topic if our domain-specifc KB or DBpedia has facts related to the keyword (see Figure 3).

3. One can choose the *Analyze* button for switching the mode to demonstrate how each sentence in the report is generated (see Figure 4). In this mode, the system shows a list of KB facts and sentence generation rules that are actually used in generating each sentence.



3D프린터

3D프린터는 3차원물체를 출력하는 프린터이다.

3D프린터는 도면인 3D도면을 이용한다.

3D프린터는 FDM, RP, SLA와 SLS를 사용한다.

FDM는 적층가공기법이다.

FDM는 ABS와 PLA를 이용한다.

SLA는 액체소재를 사용하는 적층가공기법이다.

SLS는 레이저를 사용하는 적층가공기법이다.

3D프린터는 건축, 설계와 조형물 출력에 활용된다.

3D프린터는 자동차인 스트라티를 제작한다.

스트라티는 신속성과 친환경성이 좋다.

스트라티는 무게가 200kg이다.

Figure 4: The interface of demonstration mode. Each sentence is selectable for showing all information to generate the sentence.

3.1 Report Presentation

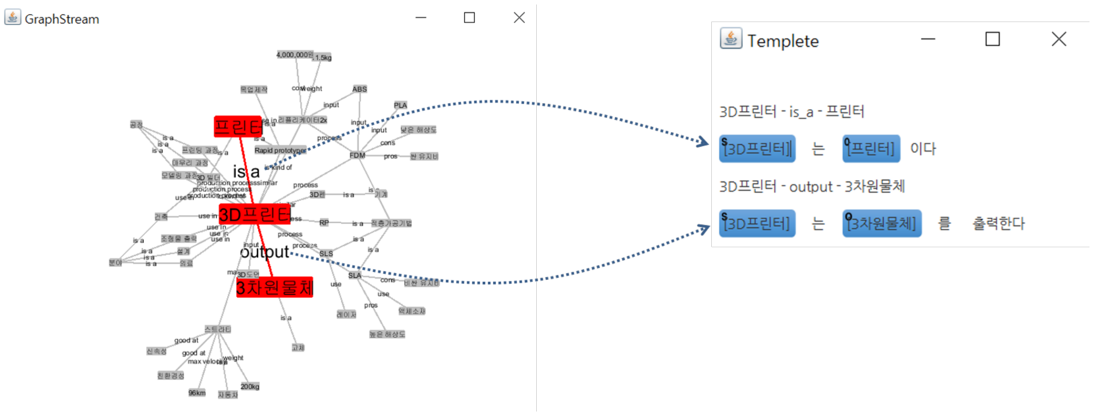
WiseReporter starts with the report cover interface as shown in Figure 2. This interface accepts a keyword as the title of a report. That is, WiseReporter generates a report about the given keyword. The sample report in Figure 3 is about ‘3D printer’. This report consists of two pages with a number of paragraphs. It provides various information about ‘3D printer’ including its definition, resources, various mechanisms, and its pros and cons.

3.2 Demonstration of Sentence Generation

Figure 4 shows the demonstration mode that is activated by choosing the *Analyze* button located on top right of the main report page. This mode lists the sentences that appear in the report line by line. Each sentence line can be selectable to demonstrate how the sentence is generated. The information provided is (i) the KB facts involved in the sentence, (ii) the templates for the facts, and (iii) the aggregation rules of the templates.

The system also shows the used KB facts as a graph (see Figure 5). The facts involved in generating the selected sentence are easily recognized as a red-colored part of the graph. The corresponding templates to the facts are also presented in Figure 5. These templates shows a notable SOV word-order characteristic of Korean well.

Finally, we can see how WiseReporter forms a complex sentence (“3D printer is a printer that prints 3D objects.”) from multiple single sentences in Figure 6. The example of Figure 6 demonstrates one of the rules combining two sentences which share a common subject. The rule in this figure



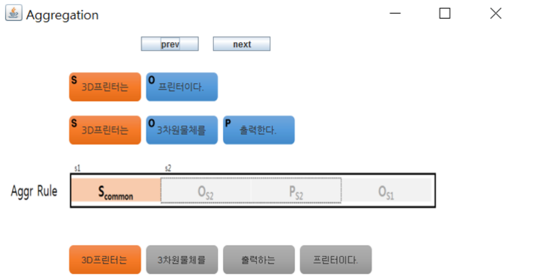
3D프린터 - is_a - 프린터

S 3D프린터 는 O 프린터 이다

3D프린터 - output - 3차원물체

S 3D프린터 는 O 3차원물체 를 출력한다

Figure 5: The graph presentation of the IT product KB focused on the topic of ‘3D printer’ (left side). The nodes and edges marked with red color are the facts used to generate the first sentence “3D printer is a printer that prints 3D objects”. The corresponding Korean templates to the facts are also presented on the right side.



S 3D프린터는 O 프린터이다.

S 3D프린터는 O 3차원물체를 P 출력한다.

Aggr Rule: s1 S_{common} s2 O_{S2} P_{S2} O_{S1}

3D프린터는 3차원물체를 출력하는 프린터이다.

Figure 6: The demonstration screen capture of the sentence aggregation. Two templates are combined into a complex sentence template.

explains how a sentence (“3D printer is a printer.”) embeds another sentence (“3D printer prints 3D objects.”) when they share a subject.

4 Conclusion

In this paper, we briefly introduced WiseReporter, a prototype text report generation system, and our demonstration of the system at IJCNLP 2017. The system uses two KBs for getting information of topics of interest, and verbalizes the information coherently. We are planning to extend it in the future by going beyond KB verbalization. We also plan to make the system available to the public by transplanting the system as a web-based service.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01772, Development of QA systems for Video Story Understanding to pass the Video Turing Test) and (No.2016-0-00145, Smart Summary Report Generation from Big Data Related to a Topic)

References

I. Androutsopoulos, G. Lampouras, and D. Galanis. 2013. Generating natural language descriptions from OWL ontologies: the naturalowl system. *Journal of Artificial Intelligence Research*, 48:671–715.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka Jr, and T. Mitchell. 2010. Toward an architecture for Never-Ending Language Learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 1306–1313.

D. Gerber and A.-C. Ngomo. 2012. Extracting multilingual natural-language patterns for RDF predicates. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management*, pages 87–96.

C. Mellish, D. Scott, L. Cahill, D. Paiva, R. Evans, and M. Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12(1):1–34.

B.-A. Nadjet, G. Casamayor, and L. Wanner. 2014. Natural language generation in the context of the semantic web. *Semantic Web*, 5(6):493–513.

P. Over, H. Dang, and D. Harman. 2007. DUC in context. *Information Processing & Management*, 43(6):1506–1520.

W.-J. Yang. 1995. *Korean language generation in an interlingua-based speech translation system*. Ph.D. thesis, Massachusetts Institute of Technology.