Korean NLP2RDF Resources

YoungGyun Hahm[1] *KyungtaeLim*[1] *YoonYongun*[2]
Jungyeul Park[3] *Key − Sun Choi*[1,2]
(1) Division of Web Science and Technology, KAIST, Daejeon, South Korea
(2) Departmentt of Computer Science, KAIST, Daejeon, South Korea
(3) Les Editions an Amzer Vak, Lannion, France
[1,2]{hahmyg, kyungtaelim, yoon,
kschoi}@kaist.ac.kr [3]park@amzer-vak.fr

Abstract

The aim of Linked Open Data (LOD) is to improve information management and integration by enhancing accessibility to the existing various forms of open data. The goal of this paper is to make Korean resources linkable entities. By using NLP tools, which are suggested in this paper, Korean texts are converted to RDF resources and they can be connected with other RDF triples. It is worth noticing that to the best of our knowledge there is a few of publicy available Korean NLP tools. For this reason, the Korean NLP platform presented here will be available as open source. And it is shown in this paper that the result of this NLP platform can be used as Linked Data entities.

Keywords: Korean Natural Language Processing, NLP2RDF, Linked Open Data.

Proceedings of the 10th Workshop on Asian Language Resources, pages 1–10,
COLING 2012, Mumbai, December 2012.

1 Introduction

Research on Linked Open Data (LOD)[1] on the Web is relatively new, but it is rapidly growing nowadays. The aim of LOD is to improve information management and integration by enhancing accessibility to the existing various formats of open data. To ease the integration of data from different sources, it is desirable to use standards (Bizer et al., 2009) such as the W3C Resource Description Framework (RDF).

There is a huge amount of unstructured text in many languages in web pages. Traditionally, these web pages have been interlinked using *hyperlinks*. However, researchers in the domain of the semantic web are focusing on data and resources, rather than web pages. In the context of semantic web, such resources are usually modelled as RDF triples (Bauer and Kaltenböck, 2011).

This paper aims to describe an NLP platform presented in (Rezk et al., 2012), but focuses on the Korean language processing. Such detailed desciption was missing in (Rezk et al., 2012). In (Rezk et al., 2012) the authors present a novel framework to acquire entities from unstructured Korean text and describe them as RDF resoruces. The main contributions of this paper are as follows: *(1)* Describing in detail how to build an open Korean NLP platform which produces POS tag, CFG and DG parsing results from one-time input; and *(2)* Providing further details on how to convert NLP outputs to the RDF. The goals of this converting are to achieve universal interoperability between the results of several NLP tools, and make Korean resource to linkable entities.

Existing Korean NLP tools, such as a morphological analyser and a syntactic parser, are reused and merged. The Sejong corpus and its POS tagset (Korean Language Institue, 2012) are used as training data. In this case the output provides RDF so entities which tokenized morpheme units have an identifier URI and can be link with existing RDF stores from the LOD-cloud. Especially, entities can be mapped with subjects in DBpedia triples.

Section 2 surveys previous work on Korean NLP and linked data. The Korean NLP platform is described with a more detailed explanation in section 3. Section 4 provides some new details on how to convert the NLP output to RDF and how to link entities with Wikipedia pages, and some tries to link entities with Wikipedia page. We discuss a conclusion in Section 5.

2 Related Work

A prime example of an NLP platform which put out RDF outputs for linked data is Stanford Core-NLP[2]. Stanford Core-NLP puts out various NLP analysis results like POS tagging, CFG parsing, DG parsing and so on for one-time input. And, by using wrapper[3] which implemented by the NLP2RDF[4] project team, those results are converted to RDF in compliance with NIF.

Actually, sharing results in Korean NLP fields is still in its early stage. Researches on Korean parsers have been focused on DG parsing *e.g.* (Chung, 2004) because Korean word order is relatively free compared to other languages. Phrase-structured Sejong Treebank is transformed into the form of DG in (Choi and Palmer, 2011). Research for CFG parsing by using Sejong Treebank has progressed (Choi et al., 2012), but it is not active and disclosure of its research results and it also true that the lack of interoperability because a variety of tools put out different format results.

In English the minimal unit for parsing is a word, but, in Korean, eojeol is a basic space unit

[1] http://lod2.eu
[2] http://nlp.stanford.edu/software/corenlp.shtml
[3] http://nlp2rdf.org/implementations/stanford-corenlp
[4] http://nlp2rdf.org

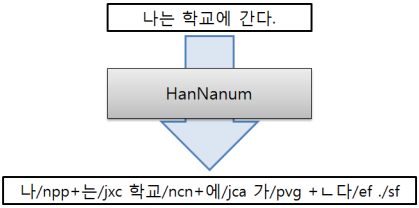

Figure 1: An example morpheme analysing Korean sentences by using HanNanum

which separated from another eojeol with white-space. An eojeol is a word or its variant word form agglutinated with grammatical affixes, and eojeols are separated by white space as in English written texts (Choi et al., 2011). Each morpheme is represented by its own POS tag so the morphological analyser is required as pre-processing for the parser. There are some existing researches about that issue and a few tools are opened already such as HanNanum (Park et al., 2010).

Research for Link Discovery issue is still on-going and there are some results such as LIMES[5] and DBpedia Spotlight[6]. Out research, such as the study of flows, attempts to outline an alternative reading of the link discovery issue. Inspiring entities are converted to RDF triples which have an URI by using our NLP platform; there are also some attempts to make links for these triples with Wikipedia page.

3 Korean Natural Language Processing Platform

Various results formats from NLP tools cause obstructive problems. So there are needs to implement one platform get one-time input. This paper describes efforts to make Korean NLP platform, and make them available as open sources. An existing morphological analyser and a syntactic parser are used and integrated in this way. Since some deficiency has been brought up, further improvement will be conducted.

And the goal of this NLP platform in this paper is that extracting entities and finding the relation between each entity from Korean resources. Morphological analyser and parser is used for this work. Details are explained in follow subsections and section 4.

3.1 Morphological Analyser

The Korean parser presented in section 3.2 requires morphologically tokenized sentences as its input. For example, English words separated by whitespace are minimal analysis units. A Korean space unit eojeol is combined with multiple morphemes. So, morphological analyser is required for splitting these morhemes from eojeol. There are two reasons: 1) most parsers consider a word which are separated by white-space as the unit of parsing. 2) For our goal, acquiring entities from Korean text. By this work, noun-tagged words can be splitted with grammatical affixes so that each word can be entities which some stacks of LOD-cloud.

As an element of Korean NLP2RDF resources, a morphological analyser, HanNanum[7] developed

[5] http://aksw.org/projects/limes
[6] http://dbpedia.org/spotlight
[7] http://sourceforge.net/projects/hannanum

```
Java –jar Berkeleyparser_korV2.jar "나는 학교에 간다."

(ROOT
        (S (NP_SBJ (NP 나) (JX 는))
          (VP (NP_AJT (NNG 학교) (JKB 에))
              (VP (VV가) (EF ㄴ다) (SF .)))))
```

Figure 2: An example phrese structured output

"나는 학교에 간다."

1.	나/NP+ 는/JX	3	NP_SBJ	
2.	학교/NNG + 에/JKB	3	NP_AJT	
3.	가/VV + ㄴ다/EF + ./SF	0	VP	ROOT

Figure 3: An example DG results

by KAIST Semantic Web Research Center[8] is employed. HanNanum was developed in C in 1999, and re-implemented in Java in 2010. It is an NLP tool which can be used independently, and include a POS tagger. HanNanum is divided into three parts depending on the level of analysis.

1. Pre-processing: Sentence boundary recognition, filtering, auto-spacing and stemming.
2. Morphological Analysis: Generate all possible morphological analysis results from each eojeol.
3. POS tagging: Assign POS tags by selecting the most probabilistic results.

HanNanum has two problems as a morphological analyser: 1) dated dictionary, 2) excessive analysis for *josa* ('postpostion'). The dated dictionary causes a lot of unknown words and might assign wrong POS tags. It amplifies eventual parsing errors. Too finely analysed josa can reate unnecessary ambiguities.

3.2 Phrase structure parser

(Choi et al., 2012) experiments with existing parsers (Stanford, Berkeley and Bikel) using the Sejong Treebank, and found out that the Berkeley parser yields the best performance. For this work, pre-processing and transforming Sejong Treebank research go on in a parallel way. Morphological analysis results from HanNanum are used as input to Korean Berkeley parser. Resource for HanNanum and the Korean Berkeley parser is open and available at the web page[9] . Specially, input system is modified for user convenience. Figure 2 shows input and phrase structured output of Korean Berkeley parser, and converted into DG results is shown Figure 3.

3.3 DG parser

There is no available Korean corpora which DG parsing can be trained. For this reason, using algorithms from (Choi, 2010), we develop a tool which converts CFG parsing results to the DG

[8] http://semanticweb.kaist.ac.kr
[9] http://semanticweb.kaist.ac.kr/home/index.php/KoreanParser

format. Our Korean NLP platform can return DG results from the CFG result. For example, Figure 3 "나/NP+는/JX" is an NP_SBJ (subject noun phrase) and dependent on "3. 가/VV+ㅏ다/EF+./SF" Our overall performance is about 72% on F1-score and it remains future work to improve performance. DG parser results show the relation between each entity in Korean texts.

4 Converting NLP outputs to RDF

Meaning of natural language processing results by converting to RDF is two-fold: 1) Universal Interoperability can be ensured, and 2) entities which be acquired from NLP tools can be linkable with LOD-cloud. This work can be able by using string URI in RDF and details are elaborate in 4.1.2.

The Korean NLP result format is different from NLP tool result format for other languages with the point of view in structure and vocabulary. To solve this heterogeneity problem, RDF is used to describe meta-data. It could be the basis of interoperability for NLP tools.

Also, for the semantic web, efforts to link RDF triples with the LOD-cloud are explained in this paper, by converting data from the web to RDF. Entities with *document level* URI can be linked with DBpedia entities with *conceptual level* URI.

4.1 Universal Interoperability

This section summarizes the NLP2RDF system presented in (Rezk et al., 2012). We focus on the methodology of the system to describe its resources. The results of the NLP tools depend on the used POS tag set, training set and their applications. To resolve this heterogeneity, discussion for NIF (Hellmann et al., 2012; Rizzo et al., 2012) is underway in NLP2RDF as a sub-project of LOD2. NIF suggests a standard for several different NLP outputs. Korean NLP results are also different from other NLP application results based on other languages. So these metadata (for the results of NLP tools) are described by using interoperable RDF. Ontology for Korean POS tags is defined, and the whole process is complying with the specifications of NIF. The results from Section 3 are converted to RDF triples.

4.1.1 Ontology for Korean Linguistic Annotations

The Ontologies of Linguistic Annotation (OLiA) (Chiarcos, 2012) are used to describe POS tags, which are different between languages. In Korean, the Sejong POS tagset and the KAIST POS tagset (Choi et al., 1994) are used for POS tagging. OLiA are offering an annotation model for Penn tag set[10] which is mainly used in English.

In the Penn tagset annotation model, POS tag information is the subpart structure of Linguistic Annotation domain. And OLiA are offering linking model also which is mapped between Penn annotation model and OLiA reference model. In this linking model, OLiA information is the subpart structure of Linguistic Concept domain and it is mapped with the POS tag set information. We made Sejong tagset anotation model and linking model which mapped into OLiA reference model for universal interoperability. And this models are posted at webpage[11]. Figure 4 shows the correspondence between Sejong POS tags and concepts in the OLiA reference model. The KAIST POS tagset will be also interoperable in our future work.

[10] http://nachhalt.sfb632.uni-potsdam.de/owl/penn.owl
[11] http://semanticweb.kaist.ac.kr/nlp2rdf/resource/

Tag		Sejong	OLiA
Super class		*LinguisticAnnotation/Tag/*	*LinguisticConcept/MorphosyntacticCategory/*
Adverb	MAJ	Adverb/ConjunctiveAdverb	Adverb and Conjunction/CoordinationgConjunction
	MAG	Adverb/GeneralAdverb	Adverb
Noun	NNB, NNG	Noun/CommonNoun	Noun/CommonNoun
	NNP	Noun/ProperNoun	Noun/ProperNoun
	NA, NF	Noun/LikelyNoun	Noun
NP		Pronoun	PronounOrDeterminer/Pronoun
Verb	VA	Verb/Adjective	Adjective/PredicativeAdjective
	VX	Verb/AuxiliaryPredicate	Verb/AuxiliaryVerb
	VC, VCN, VCP	Verb/Copula	Verb
	VV	Verb/VerbalPredicate	Verb
	NV	Verb	Verb
SN, XN		CardinalNumber	Quantifier/Numeral
MM		Determiner	PronounOrDeterminer/Determiner
SH, SL		ForeignWord	Residual/Foreign
IC		Interjection	Interjection
SE, SF, SO, SP, SS		Symbol	Punctuation
superclass		*LinguisticAnnotation/Tag/*	*MorphologicalCategory/*
Particle	JC, JX	Particle/AuxiliaryPostposition	Morpheme/MorphologicalParticle
	JKB, JKC, JKG, JKO, JKQ, JKS, JKV	Particle/CaseMarker	Morpheme/MorphologicalParticle
	XPN	Particle/Prefix	Morpheme/MorphologicalParticle/prefix
	XSA, XSN, XSV	Particle/Suffix	Morpheme/MorphologicalParticle/suffix
	EC, EF, EP, ETM, ETN	Particle/VerbalEnding	Morpheme/MorphologicalParticle/suffix
	XR	Particle/Radical *(Mapping with LikelyNoun)*	Morpheme

Figure 4: Correspondence between Sejong tags and concepts in OLiA

4.1.2 String URI

The advantage of specifying URIs for each entity (results of NLP) is two-fold: 1) entities can be described with RDF, and 2) entities can be linked-able with other RDF triples. Our goal is not only focusing on getting universal interoperability for NLP output, but also making links with LOD-cloud for Korean resources from the text. Therefore, specifying an URI for each entity is an important task.

Each noun is a morpheme unit, so URIs specification works on morpheme units. Recognized entities by morphological analysing are URI specified so that it is used as 'Subject' in RDF triples. And morpheme unit entities, especially nouns, can be used as some stacks for LOD-cloud. This will be explained in Section 4.2.

NIF provides two URI schemes: The offset and context-hash based schemes. The Hash-based URI is used for our Korean NLP platform results.

Figure 5 shows an example about URI specification for Korean word.

"나는 학교에 간다" (I go to school)

'학교' (school)
Hash_15_2_c3508b1509ed7789297de77cfd9fb14f_학교

Figure 5: An example URI specification for Korean word

4.2 Linking Korean Resources with the LOD-cloud

Each of the morphemes obtained by NLP tools appears in RDF as entities with a POS tag and an URI. Actually this URI is at *document level*. Entities which appear in sentences have limited meaning in a sentence, document, text, or web page, and the described RDF information is just POS and grammatical role. These entities are isolated and restrictive.

Producing RDF from NLP results makes entities get an URI so they can be linkable. With these URIs we can link entities with RDF resources LOD-cloud, for example, DBpedia. We assume that entities of DBpedia have *conceptual level* URI and reliable information enough as a collective intelligence.

Our goal is to link Korean resources converted to RDF by NLP tools to stacks of the LOD-cloud. For example, some system can show the grammatical role of entities and DBpedia triple information. First step for this goal is to make links for entities to the Korean Wikipedia pages. The purpose of this work is the first step in the progress of experimenting Link Discovery for Korean. Our approach first accesses NLP results which obtains all nouns. We then query the Noun class using the data property anchor of to get the string. System checks the Wikipedia page for such a string. If a page exists, a link is created for the Wikipedia page. That is why we use a Wikipedia page because DBpedia always shows always page even there are no information for strings.

This system has been implemented on the web site (Semantic Web Research Center, 2012), available to anyone. This web site provides the following two functions: 1) Return a variety NLP results such as morpological analysing, CFG and DG parsing from one input text; 2) Make Links to Korea Wikipedia pages for each entity which noun-tagged words.

Here on link discovery of Korean resource, two issues are still remaining:

1. Efficient algorithms for approaches to DBpedia.
2. Using synonym relations.

Our current link discovery method is just string matching with Levenshtein distance. This method is simple; there are many limitations to finding links this way. As many other approaches such as LIMES, we need to develop efficient and adequate an algorithm for Korean.

Using synonym relations can be an alternative way to extend the capabilities of the link discovery approach. CoreNet (Choi, 2003) will be considered as pertinent resources for this issue.

5 Conclusion and Future Work

The focus of this study is two-fold: First, we developed the Korean NLP platform which returns a variety of NLP result outputs from one-time input, and make it publicly available. However, further improvement for Korean NLP tools is required for Link discovery. In particular the morpheme

analyser should be modified. Second, we provided further details on how results of NLP tools are mapped into RDF. It is worth noticing that the RDF triples generated in this framework follow the NIF standard. In particular, the Sejong tagset linking model is built and used for universal interoperability for Korean NLP. Moreover, resolving the link discovery issue will be our future work.

Acknowledgement

This research was supported by the Industrial Technology International Cooperation Program (FT-1102, Creating Knowledge out of Interlinked Data) of MKE/KIAT.

References

Bauer, F. and Kaltenböck, M. (2011). *Linked Open Data: The Essentials*. Edition mono/monochrom, Vienna.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *Journal on Semantic Web and Information Systems (IJSWIS); Special Issue on Linked Data*, 5(3):1–22.

Chiarcos, C. (2012). Ontologies of linguistic annotation: Survey and perspectives. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Choi, D., Park, J., and Choi, K.-S. (2012). Korean treebank transformation for parser training. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 78–88, Jeju, Republic of Korea. Association for Computational Linguistics.

Choi, J. D. and Palmer, M. (2011). Statistical dependency parsing in korean: From corpus generation to automatic parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Choi, K.-S. (2003). CoreNet: Chinese-japanese-korean wordnet with shared semantic hierarchy. In *Proceedings of Natural Language Processing and Knowledge Engineering*, pages 767–770.

Choi, K.-S., Han, Y. S., Han, Y. G., and Kwon, O. W. (1994). Kaist tree bank project for korean: Present and future development. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14.

Choi, K.-S., Isahara, H., and Sun, M. (2011). *Language resource management - Word segmentation of written texts - Part 2: Word segmentation for Chinese, Japanese and Korean*, ISO 24614-2.

Choi, Jinho D.; Palmer, M. (2010). Robust constituent-to-dependency conversion for english. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT'9)*, pages 55–66, Tartu, Estonia.

Chung, H. (2004). *Statistical Korean Dependency Parsing Model based on the Surface Contextual Information*. PhD thesis, Korea University.

Hellmann, S., Lehmann, J., and Auer, S. (2012). Linked-data aware uri schemes for referencing text fragments. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW2012)*, Galway City, Ireland.

Park, S., Choi, D., Kim, E., and Choi, K.-S. (2010). A plug-in component-based korean morphological analyzer. In *Proceedings of HCLT2010*, pages 197–201.

Rezk, M., Park, J., Yoon, Y., Lim, K., Larsen, J., Hahm, Y., and Choi, K.-S. (2012). Korean Linked Data on the Web: From Text to RDF. In *Proceedings of JIST2012:Joint International Semantic Technology Conference*, Nara, Japan.

Rizzo, G., Troncy, R., Hellmann, S., and Bruemmer, M. (2012). NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW 2012, 5th Workshop on Linked Data on the Web, April 16, 2012, Lyon, France*, Lyon, FRANCE.

Semantic Web Research Center (2012). Korean NLP2RDF demo site, `http://semanticweb.kaist.ac.kr/nlp2rdf`, KAIST.