# **Brochette: Toward A Continuous Learning Platform for Knowledge Acquisition and Integration**

Ivan Berlocher, Kim Seonho, Tony Lee

Saltlux Inc, Daewoong Bldg. 689-4, Yeoksam 1 dong, Gangnam-gu, Seoul, South Korea {ivan, shkim, tony}@saltlux.com

#### Abstract

Question-Answering (QA) systems like IBM Watson are particularly challenging to design and need to cover areas including computational linguistics, information retrieval, knowledge representation and reasoning, and machine learning. 'Exobrain' is a Korean Research Program which aims at building such a high-performing QA system for the Korean Language. In this paper, we describe 'Brochette', a continuous learning platform that iteratively acquires a large volume of unstructured data and extracts sentences related to entries of an ontology, a formal representation of knowledge used for further queries and reasoning purposes. As various Machine Learning modules must be trained and tuned to suit new content and adapt to new domains, Brochette platform also provides a framework for continuously re-training and evaluating machine learning components for syntactic and semantic analysis. Integrating semantic information from the ontology and results of these machine learners, the system is able to discover new knowledge which is then incorporated again in the Knowledge Base, making it continuously evolving.

#### Introduction

We describe Brochette, a system for semi-automatically acquiring knowledge from online sources like encyclopedias or dictionaries and producing semantic annotation corpora for further various machine learners. These machine learning agents are able to discover new facts and entities, populating a very large Knowledge Base dedicated to enlarging the coverage and precision of a high performance Question-Answering System for the Korean Language.

Our work is inspired by NELL (Carlson et al. 2010), a system which acquires two types of knowledge:

(1) knowledge about which noun phrases refer to which specified semantic categories, such as cities, companies, and sports teams, and

(2) knowledge about which pairs of noun phrases satisfy which specified semantic relations, such as hasOf-ficesIn(organization, location).

We have the same goal as NELL in term of self-learning for populating a Knowledge Base (KB), but our approach is radically different. In NELL, four main components are tightly designed for learning facts about the two types of knowledge describe above, whereas our platform is a generic framework which can plug in any type of annotator. The annotators can be cascaded in customizable workflows to learn new knowledge. The Brochette ontology is expressed in RDF, whereas the NELL ontology is not.

## **Input Ontology**

The Knowledge Base is an ontology called "XB Ontology" and the schema (classes and properties) has been built manually from analysis of various resources like Wikipedia, YAGO (Suchanek, Kasneci and Weikum, 2007), DBpedia and KorLex, the Korean WordNet from Busan University (Yoon, Hwang, Lee and Kwon, 2009), which contains mapping information to the original Wordnet synonyms set. To define classes, at first, we got the intersection of KorLex nouns and YAGO classes. KorLex includes 100,000 nouns that are sufficient to describe knowledge in Korean, while YAGO classes fully cover the instances of Wikipedia. Because both YAGO and KorLex have mapping information to WordNet, calculating intersection of KorLex nouns and YAGO classes is straightforward. As a result, 5,775 classes were derived. To simplify the integration of NLP results to the ontology the Tag set of a Named Entities Recognizer (NER) was also used. Similarly as YAGO (Suchanek, Kasneci and Weikum, 2007) semi-structure information of Wikipedia,

Copyright  $\[mathbb{C}\]$  2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

particularly info-boxes and categories have been used to make rules for extracting automatically instances rules Wikipedia (the Korean version) as shown in Figure 1.

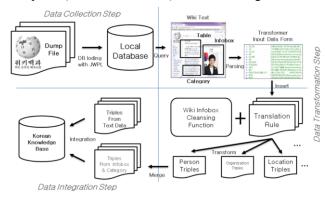


Figure 2: Automatic Instances creation from Wikipedia

For the properties, from the analysis of Wikipedia info-box attributes, YAGO properties and DBPedia properties, 129 properties have been first defined. As semi-automatically constructed ontology from Wikipedia semi-structured information has limitations both in terms of coverage and precision, human curators are needed to check the results of automatically extracted ontology, had new classes, instances or properties base on the text in Wikipedia main page. A web based semantic annotation tool allows curators to semantically annotate documents in RDFa and edit (add, remove, modify) the ontology. At the current state of this project, the XB Ontolgy has is 6,116 classes with 583 properties, 229,654 instances resulting in 1,632,861 triples.

#### **Brochette System Architecture**

One key requirement for a successful QA is to acquire high quality corpora from which new knowledge can be extracted. Despite the breadth of knowledge on the Web, experiments (Schlaefer et al., 2011) show that QA performance does not necessarily improve or may even degrade if sources are indiscriminately added. We follow their source acquisition procedure, which is an iterative development process of acquiring new collections of documents to cover salient topics. The design of the system is depicted in Figure 2.

#### **Sources Acquisition**

In the case of Watson (Schlaefer et al. 2011), acquiring general-purpose title-oriented document sources such as Wikipedia has been shown to improve QA performance, so we followed this approach for our platform.

In order to acquire knowledge we designed a Fetcher that can crawl News and Blogs from the Web in real time, a Batch Downloader that downloads sources like Wikipedia or Wiktionary and a Meta-Search engine that crawls documents from top results of various search engines like Google or Yahoo! corresponding to a given query. This component is intended to find documents related to a given entity in cases where the previous corpus does not contain enough documents. It is also used to find related or similar documents for a given document to expand the knowledge base. The documents are then indexed by a local search engine for fast retrieval of passages, sentences or documents related to a given topic.

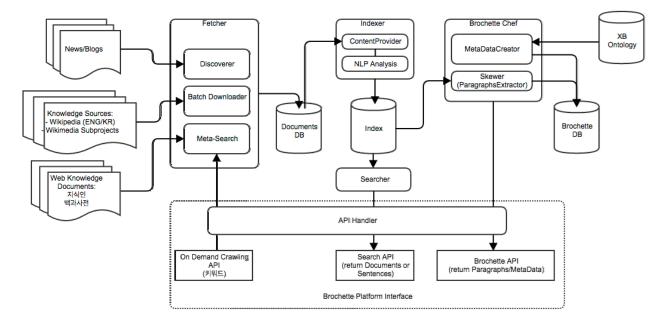


Figure 1: Brochette System Overview

### **Ontology Integration**

We designed a component called 'Brochette Chef' that produces a set of sentences semantically related to an instance of the 'XB Ontology'. The Brochette Chef module comprises 2 components: the Metadata Creator and the Skewer. The Metadata Creator iterates over each entity of the ontology and gets semantic metadata like class information (Person, Company etc.), synonyms, related entities etc. from which it makes a vector of features capturing the semantic of the entities. Our approach follows the DBpedia spotlight (Mendes et al., 2011) algorithm. The entity of the ontology has a description page related to a dictionary or encyclopedia like Wikipedia, so the full text is analyzed, removing common nouns, stopwords, and words in hyper-links, categories or info-boxes are kept. As explained by Mendes, the weighting of features uses the TF (Term Frequency) with ponderation by the ICF (Inverse Candidate Frequency). With this vector representation (Vector Space Model) of each entity of the KB, the semantic of the entity is kept so we call it the 'semantic vector' and it serves as a basis for later semantic matching of entities in text.

The skewer uses the semantic vector related to one KB entry to get the top documents related to the entry from the internal search engine then splits each document into paragraphs and sentences. For each sentence or paragraph, a similarity between the semantic vector of the ontology instance and the features vector of the sentence is calculated and only the sentences or paragraphs with a high threshold similarity are kept. The set of all sentences is stored as a pseudo-document, called a 'brochette'. The figure 3 below illustrates an example of a brochette for the entry 'Hee Yo-Ri' a famous Korean female singer.



Figure 4: Example of a brochette in JSON format

For analyzing each sentence, several Natural Language Processing (NLP) components are used, such as Part-ofspeech (POS) Tagger, Named Entities Recognizer (NER) and dependency parser. All these modules came from different organizations, are implemented in several languages like C++ or Java, have various output formats and specific settings for executing them. In such an environment, we developed a 'Continuous Learning Framework' based on UIMA (Ferrucci and Lally, 2004) for standardizing the input/output of modules using CAS, the Common Analysis Structure for the annotations produced, and UIMA-AS (Asynchronous Scaleout) for supporting distributed computing.

#### **Continuous Learning Framework**

Annotator Engines (AE) in the UIMA terminology sense, like NER or Tagger, are Machine Learning based algorithms and they should be able to be re-trained online given feedback about incorrectly tagged results.

For training annotators a training set has to be created; to be evaluated a test set has to be provided too. This is the role of the Corpus Manager, Evaluations Handler and

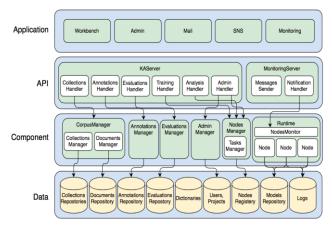


Figure 3: Continuous Learning Framework

Training Handler. The performance of each annotator has to be monitored over time assuring that the precision and recall of the overall system is increasing. This functionality is assured by the Evaluations Manager. The framework also provides a pluggable mechanism allowing adding or interchanging any other Annotator. We implemented a component called a 'Node' providing a Server/Client wrapping of the core NLP engines with a common interface incorporating four main methods: train, load, evaluate, analyze. The result of each engine is encapsulated in a CAS container provided by the UIMA framework. The framework provides a Graphical User Interface (GUI), shown in figure 5, through which human curators can easily:

TagSet	Stags								
2 Engines		TOURNON TOWNER TO	ICOND						
								C	Multi-Highligh
Evaluation	All Docum	ents 1					(	2 Search	
Online Test	Color T	Online Test - 2014-04-17 15:02:50							
and the second			이 것으로 확인된다. 로이터뷰하이 물론						
	특징용 것 는 점 말에 이드로 등 례료가 다 레도우 당 시 받아봄 레일에서	2추개 되는지는 자세히 설명되지 이다. 다비지는 로이름플러 보도기 등각하도록 설계였다. 이후 내놓은 올레 내놓을 스마트폰에 다시 안 사용한 비진이다. 지난 3월 구렬이 지된 <u>스마트폰</u> 과 연결해 함 수 있 우 수 있도록 한 것이 홈페트로이드해 일 가지 의미를 찾아보자. 지금까	1021년2, '기어맛을 내셨다. 용한길 않았다. 유아동신과 가진 엔터웨어서 나 나 및 협정문제의 다음 스마트시제 기어갔지 1021년20년 프라이팬 05가 프라테트 05가 월려간다면, 구글의 '인 처음으로 소개했다. 올해 2초가 만으로 도록 돕는다. 구글지도나 형어웃 증구길 이외 특십이다. 구글로 동생정형 기술 1031특십이다. 구글로 동생정형 기술 1031특십이다. 구글로 동생정형 기술	도 스마트시계는 월 의 안드로이드웨어가 탈채됐다. 기여빛은 드로이드웨어 일 가 LG전자와 모토로링 일 기본 서비스를 ( 도 안드로이드웨어를 당한으로 모으지 못했	게 안금됐용 뿐이 들어간다는 것입 실시간 운영제7 농성이 높다. 안5 에서 실계 제품 는목에서 활용할 탑재한 스마트 몇인 것이 아닐까	다. 윈트는 있다. 구 을 확인했다고 진행 ((RTOS) 로 움직인 트로이드웨어는 기관 을 내놓을 예정이다 수 있다. 특히 구글 에게에서 활용할 수 하가 한번 바뀌는	글 만드로이드 위 다. 삼행전자의 및 다. 운항길 수석부 한 만드로이드 아이 아드로이드웨어 나무 서비스를 함테 있을 것으로 기대 함은 시간 동안 해	안에서제(05)여 번째 스마트시 사장의 성영호 등 웨어러를 7 는 스마트시계 용해 개인화된 된다. 협렬권과 다의 제품 카이	에 기반물 둔다 시계는 안드로 한 것처럼 <b>정정</b> 네기와 잘 <b>머물</b> 볼 안드로이드 정보물 순목에 해외 이 같은 볼 데고리에서 3가
	는 태왕의 서 주유는 분위리가	[레다. 반대로 삼성전자의 5번째 만지 못했다. 만드로이드웨어의 의 비 지금과 달라질 것으로 기대하는	제품에서 시장이 표동될 가능성도 있다 미는 웨어러블 시장에 구글이 힘을 다친 까닭이다. 윤한길 주석부사장은 인터뷰	. 그동안 스마트시계 IFL는 데 있다. 05의 에서 <b>'우리</b> 는 뛰어리	황용성을 담보 분이나 홈 컨버터	로 더 많은 제조업의 빈스, 자동차 등 세5	가 뛰어들 가능성 2운 분이에 도전해	이 열렸다. 🎒 1고 있다"라며	아트시계시장
	는 태왕의 서 주유는 분위리가	IMC: 반대로 삼성전자의 5번째 2지 못했다. 만드로이드웨어의 의 지금과 달라질 것으로 기대하는 1제품 판매에 시너지효과를 가져!	제품에서 시장이 國際될 가능성도 있다 미는 웨어러블 시장에 구글이 힘을 다친 까닭이다. 문환질 <b>주석부사장은 인터</b> 뷰 을 것"이라고 답했다. 삼성권자의 다음 /	. 그동안 스마트시계 ICL는 데 있다. 05의 에서 "말립는 웨어리 스마트시계는 물해 6	I 활용성을 담보!  쓴이나 홈 컨버 I바반기 동장할 것	2 더 많은 제조업세 선스, 제동차 등 세3 으로 보인다. 聖동해	가 뛰어들 가능성 I운 분이에 도전해 [기차 sideway@	이 영랐다. 전 1고 있다"라며 Roloter.net 41 pro	<u>마트시계 시</u> 용 '장기적으로 tv next 바
	는 태현감 시 구석문 평양[77] 프라이엄 ************************************	INC: 반대로 삼성전처의 5번째 문지 못했다. 만드로이드웨어의 의 처금과 당하철 것으로 기대하는 계종 판매에 시나치효과를 가져 (V_PRE(I)) © PE_MYP(I) © ME(I) 한 ICE_MOLOGIAN(I) 한 PE_	제품에서 시장이 표동될 가능성도 있다 미는 웨어러블 시장에 구글이 힘을 다친 까닭이다. 윤한길 주석부사장은 인터뷰	. 그동안 스마트시계 (다는 대 있다. 059 에서 "문력는 웨어리 스마트시계는 몸해 6 SDCHTHSI & DEG 3	I 활용성용 당보 변한이나 홈 컨버터 아반기 동강할 것 POINS(2) ♥ 056, PARTID ♥ AU,IEO	E 미 양은 제조업의 변소, 제동차 등 세5 으로 보인다. 문문해 MLFART(1) 06,0 RD(1) 등 05,75AR(2)	가 뛰어들 가능성 1은 선이에 도전해 12차 sideway@ D4RS(1)* TM_Sid 0 (1,1555H(1)*	ICP_COUNTRY	대표시계 시용 *경기적으로 tv next \+ (2) % cv_Aducri (3) % qt_count
	는 태현감 시 구석문 평양[77] 프라이엄 ************************************	해지는 가수도와 상성원과의 SPR해 문자 못했다. 만드로 상성원과의 SPR해 지금과 달라철 것으로 기구하는 제품 판매에 시네지효과를 가져! © CV_PRECUS® PL_MYDEDS® MRC_DREMECK_ID ® CV_PORTO	제품에서 시장이 國務될 가능성도 있다 미는 웨어러블 사장에 구름이 항을 다른 까당이다. 물란질 출석별사물은 인터블 후 것'이라고 당했다. 삼성원자의 다음 / P5_UMAR(1) % OSG, MUSAES(1) % OSG SOCAL_SCENCE(1) % CV_388(2) #49AA	. 그동안 스마트시계 (다는 대 있다. 059 에서 "문력는 웨어리 스마트시계는 몸해 6 SDCHTHSI & DEG 3	I 활용성용 당보 변한이나 홈 컨버터 아반기 동강할 것 POINS(2) ♥ 056, PARTID ♥ AU,IEO	E 미 양은 제조업의 변소, 제동차 등 세5 으로 보인다. 문문해 MLFART(1) 06,0 RD(1) 등 05,75AR(2)	가 뛰어들 가능성 1은 선이에 도전해 12차 sideway@ D4RS(1)* TM_Sid 0 (1,1555H(1)*	ICP_COUNTRY	전프시계 시공 *경기적으로 tv next ₩ (2) % CV_FOLICH (5) % 0T_COLNT MCCHL_TISUE % LNAYCHL
	EB283 A 4 44 BERZ 71 BER 019 A Annot	BBCL (FICER 상정원자의 SPER EV 못했다. 안드로이드웨어의 의 지금과 달라철 것으로 기다하는 제품 판매에 시네지효과를 가져? © CUJNETUL® PLANTHOL® BBCTI® LEG.MOLADAMTI® @ BBCTI® LEG.MOLADAMTI® DE BBCTI® LEG.MOLADAMTI® DE BBCT	наран Аропавар тыраба стра оправлять правар тыраба стра оправлять правар така и правалять правар с права и права с права с права и права с права и права с права с права с права с права и права с права с права с права с права с права и права с права с права с права с права с права с права и права с права	그동안 스마트시키 FI는 데 있다. OSS 에서 "해외는 웨어라 스마트시키는 동네 6 SOCETTISI © OGG, 3 SOCETTISI © OGG, 4 SOCETTISI © OGG, 4 SOCETISI ©	1 활용성을 당보낸 철미나 홈 컨터뷰 바인기 동장할 것 POETS(2) ♥ DSG, PAETS(2) ♥ DSG, PAETS(2) ♥ ASA, PAETS(2) ♥ ASA,	은 더 많은 제조업의 선소, 작동작 등 세려 으로 보인다. 활동작 MLIPART() © 06,0 AD(1) ® 07,9FAR(1) © 07,	2) 밖이를 가능성 1은 산이에 도전해 (기자 Sciewayd) (기자 Sciewayd) () () () () () () () () () () () ()	IO STATE STATE	대표시계 시장 *장기려으로 tv next ₩ (2) ♥ CV_POLICY (5) ♥ CV_POLICY (5) ♥ CT_COLM ♥ CRU_19508 ♥ CRU_19508
	EB283 A 4 44 BERZ 71 BER 019 A Annot	BRC: 반영감 상상유가의 Steel 지금과 달라질 것으로 기대해는 지금과 달라질 것으로 기대해는 또 CL/NRICH 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED Text 5	(Кран А. (20) (総要選う) うちない (20) (20) (20) (20) (20) (20) (20) (20)	그동안 소마트시키 (다는 데 있다, 052 에서 개별된는 데이츠 스마트시키는 몸에 이 스마트시키는 몸에 이 스마트시키는 몸에 이 스마트시키는 몸에 이 스마트시키는 몸에 이 스마트시키는 위 스마트시키는 위 스마트시키는 위 스마트시키는 위 스마트시키는 이 스마트시키는 이 스마트 스마트시키는 이 스마트 스마트시키는 이 스마트 스마트 스마트시키는 이 스마트 스마트시 스마트 스마트 스마트 스마트 스마트 스마트 스마트 스마트 스마트 스마트	1 활용성을 당보년 철미나 홈 컨터뷰 바인기 동장할 것 POETS(2) & DSG, PARTID & AV, RO PARTID & AV, RO PARTID & AV, RO PARTID & AV, RO	은 더 많은 제조업에 선소, 제동과 등 45 으로 보인다. 모등에 MERANY(1) 이 06,0 AC(1) 이 05,75A(0) URMATARIES 이 07,	가 밖이들 가능성 또 선이에 도전해 17년 sidewayd nerocu 등 740,58 이 대,LINSHU 등 JACON E LCP,C	ICI STATE	대표시계 시장 '장기격으로 tv next ዙ (2) 는 cv_Poucy (5) 는 qt_court HM_COUL_TINUS 는 LINDOW
	는 태환파 시 주 사는 해외2가 제외0명 • AMJ · · · · · · · · · · · · · · · · · · ·	BRC: 반영감 상상유가의 Steel 지금과 달라질 것으로 기대해는 지금과 달라질 것으로 기대해는 또 CL/NRICH 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED Text 5	부분(Rote AL 2010) 표준철 가능성 또 있다. 이는 목이가관을 사용하고 중에서 구금이 있는 이는 가능(Rote) 운동감고 중에서 부분(Al 2010) 구금이 가능(Rote) 운동감고 중에서 부분(Al 2010) 구름 2 210(H2) 도 당첨도 스럽 성원자(이 다음 / PL_(NAME(1)) 등 OCE_(LUNARS(K)) 등 OCE SCOR_COLOCION(ONU) 한 (LONARS(K)) 등 OCE SCOR_COLOCION(ONU) 한 (LONARS(K)) 지하 (COLOCION(ONU) 한 (LONARS(K)) 등 OCE SCOR_COLOCION(ONU) 한 (LONARS(K)) 지하 (COLOCION(ONU) 한 (LONARS(K)) 등 OCE SCOR_COLOCION(ONU) 한 (LONARS(K)) 등 OCE (LONARS(K)) 등 OCE_(LONARS(K)) 등 OCE SCOR_COLOCION(ONU) 한 (LONARS(K)) 등 OCE SCOR_COLOCION(ONU) 한 (LONARS(K)) 등 OCE (LONARS(K)) 등 OCE_(LONARS(K)) 등 OCE SCOR_COLOCION(ONU) 한 (LONARS(K)) 등 OCE SCOR_COLOCION(ONU) E (LONARS(K)) = OCE SCOR_COLOCION(ONU) E (LONARS(K)) = OCE SCOR_COLOCION(ONU) E (LONARS(K)) = OCE SCOR_COLOCION(ONU) = OCE SCOR_COLOC	그 동안 스마트시키 ICH는 데 있다. (53 스마트시키는 웨어식 스마트시키는 음력 6 SOCETYISI © GGL 3 ART_CRAFT(I) © AM, ( 한 RO,MI(I) 한 R) Start Offset 1541	환용성을 당보 플라나 홈 컨테 바반기 동강할 것 PORTICI & ACE, PARTICI & AL, RO PARTICI & AVAR, End Offset 1566	은 더 많은 제 조업에 선스, 제동과 등 45 으로 세인다. 모등해 MERAIN(1) 등 06,0 AD(1) 등 07,17AM(1) UTRATURED 등 07, Created Date 2014-04-17	가 뛰어들 가능성 또 산이에 도전용 (기자 sidewayd netRict) 는 TAUSA 이 다.tRISHU 는 TAUSA (Attended To LCP, C Annocator admin	ICI STOLE 20 2 SCIPTION BIODEFINET ICP, COUNTRY ANTAL CIVID TO Status Status Status	(1111년 1111년 11111년 111111
	는 명령원	BRC: 반영감 상상유가의 Steel 지금과 달라질 것으로 기대해는 지금과 달라질 것으로 기대해는 또 CL/NRICH 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED Text 5	New York New York	그동안 소마트시키 (다는 데 있다, 059 소마트시키는 웨이츠 소마트시키는 홈페 6 500EFY(1) 등 066,5 (APE, DAPT(1) 등 APA, APE, DAPT(1) 등 APA, APA, DAPT(1) \in APA, DAPT(1) \in	환용성을 당보 봄이나 총 전매 반인기 동장할 것 PORTIO & OSG, PARTIO & Ar JRO PARTIO AR	E 디 당은 제조업체 선스, 제동과 등 45 으로 보인다. 모름해 MLEART(1) 등 06,0 DE(1) 등 05,114A(1) EPERATURE (1) 등 01, Created Date 2014-04-17 2014-04-17	가 뛰어들 가능성 은 선어에 도전을 (112) sidewayd merscl) 등 netca grusson(1) 등 daterol 등 LCPUC Annotator admin admin	이 영었다. 20 12. 있다가라이 Blober met et pre top, country AntiALCIVID 한 Status 성명 성명	대표시계 시장 '장기격으로 tv next ዙ (2) 는 cv_Poucy (5) 는 qt_court HM_COUL_TINUS 는 LINDOW
	는 대학교 시 구 국왕 동태(2) 가 제고(1) 함 · AMJ 이 전 Annot 이 면서 가지 및 이 전 20 가 제고(1) 한 · · · · · · · · · · · · · · · · · · ·	BRC: 반영감 상상유가의 Steel 지금과 달라질 것으로 기대해는 지금과 달라질 것으로 기대해는 또 CL/NRICH 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED MERT 한 ERLANDER 한 PLANDED Text 5		그 문한 스마트시가 2 - 고환한 스마트시가 2 에서 1월문 030 스마트시가는 문제 0 스마트시가는 문제 0 	환용성을 당보 봄이나 홈 전매 바반기 동상할 것 POETICS & OSA, PARTIS & AJ, PO AMICO AMIA End Offset 1506 1509 1475	전 이 양은 제 조업체 선스, 제동과 등 443 으로 체인(C), 양동형 ALL(DAM(L)) 등 00,0 AD(L) 등 01,912AD(L) (Created Date 2014-04-17 2014-04-17	가 뛰어들 가능성 은 선어에 도전을 (기가 sidewayd) nerscl) 등 Twijse 이 gijsecht 등 try, c Anecosor admin admin admin	이 영었다. 월 13 있다가라이 Blobermit 44 pro 10 한 Taujuw 10 CP, CDuster were CPV 20 한 Status 444 444	(1111년 1111년 11111년 111111
	는 대환권 시 주 수정 비료[27]가 비료[01] © Antol (19] (19] (19] (19] (19] (19] (19] (19]	BED: 반신의 상상문가의 정당된 기공급 및 당각철 성으로 기대하는 기공급 당각철 성으로 기대하는 계공 단역에 서비치, 호·미, 기대 등 CUPREID 등 CUPREID 등 CUPREID 등 CUPREID 등 CERNICATIO 등 CUPREID 등 MCONNECATIO 등 CUPREID ation Information Text 5 addressy@Bioter.net	HERHAN (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980) BERRER 716 (1980)   US MORINA (1980) BERRER 716 (1980)	그 문란 스마트시가 다는 데 있다. 059 이시 "행관는 에이스 스마트시키는 웨이스 스마트시키는 웨이스 SOCETHSI © 006.5 AAC_CRATICI © AAL & CRATICI © AAL & CRATICI © AAL STATE Officet 1541 1555 1479 1439	환용성용 당보: 생리나 총 전매 바반기 동장할 것 postsc3 © odds, patrici © Adjico patrici © A	은 더 같은 제조업의 변스, 제품과 등 45 으로 M인다. 호텔 MLEART() 등 00,0 AD(1) 등 02,14A(1) UREATURE() 등 01, Created Date 2014-04-17 2014-04-17 2014-04-17	2) 뛰어들 가능성 요즘 성이에 도전용 212 Sideway® 마타RCU 등 TALSA 이 다.ISSS (C.P.C Annotator admin admin admin admin	이 영었다. 실 ILD 있다가라며 Rober met et pre- et CP, CDUMRY ANTIAL CIVID ® Status 성명 영명 영명	대표시계 시장 '장기격으로 tv next ዙ (2) 는 cv_Poucy (5) 는 qt_court HM_COUL_TINUS 는 LINDOW
	는 대왕권 시 주 우성 동동(27)가 동조(10)당 이 Annot (11) (12) (12) (12) (12) (12) (12) (12)	BED: 반신의 상상문가의 정당된 기공급 및 당각철 성으로 기대하는 기공급 당각철 성으로 기대하는 계공 단역에 서비치, 호·미, 기대 등 CUPREID 등 CUPREID 등 CUPREID 등 CUPREID 등 CERNICATIO 등 CUPREID 등 MCONNECATIO 등 CUPREID ation Information Text 5 addressy@Bioter.net		그 문한 스마트시가 다는 데 있다. OS의 이시 1월3년 - 베이스 스마트시가는 웨어스 - SOUTHYN & GEG. - SOUTHYN & GEG. - SOUTHYN & AM. - 한 HOLANTIN & AM. - 한 HOLANTIN & AM. - 1541 - 1505 - 1420 - 1420 - 1420	활용성을 당보 분량이나 총 전력( 바반기 동강할 것 AA(10)	은 다 같은 제조업의 전소, 제공자 등 445 으로 세인다. 호텔 MLEART(1) 06,0 BC(1) % DT, TARCO DT, TARCO DT, TARCO Created Date 2014-04-17 2014-04-17 2014-04-17	71 데이플 가능성 요즘 분야이 도전해 로구 분야이 도전해 로구자 Sciences 전 이 도자 Sciences 전 이 도자 Sciences 전 Annocator admin admin admin admin admin	이 영었다. 20 13 있다가라며 Robernat: 44 pro- 10 등 14 pro- 10 Fro- 10 F	대표시계 시장 *장기려으로 tv next ₩ (2) ♥ CV_POLICY (5) ♥ CV_POLICY (5) ♥ CT_COLM ♥ CRU_19508 ♥ CRU_19508
	는 대통령권 시 구 유식 통령권가/ 문리(11) (11) (11) (11) (11) (11) (11) (11)	BBL: 한국내 삼성전위, BBL 가장과 영소는 1952-0553(Rel B) 가장과 가장과 영소한 2023 가격하는 가장과 영소한 2023 가격하는 가장과 Rel Rel Rel Rel Rel Rel BBC 10 102, Rel Rel Rel Rel BBC 10 102, Rel Rel Rel Rel BBC 10 102, Rel Rel Rel BBC 102, Rel Rel Rel BBC 102, Rel Rel Rel Rel Rel Rel BBC 102, Rel Rel Rel Rel Rel Rel Rel Rel Rel BBC 102, Rel	Heijerk (1980) BBBBB 716 (555 GE)   USKNIDE 1000 (1990) Tele 100 (1990)   Miglicita Libert (1990) Statistical USKNID   Miglicita Libert (1990) Statistical USKNID   DODU, STATISTIC (1990) Statistical USKNID   DODU, STATISTIC (1990) Statistical USKNID   NUM, ANT, AND COLLECTION (1990) Statistical USKNID   Tele OV, 20020   NUM, ANT, CANPY T PL, MAT   MIG, MAT, CANPY T PL, MAT   MIG, MAT, CANPY T MAN, MAT, CANPY T	. 교육한 스마트시키 는타는 데 있는다. OSS 니다는 데 있는다. OSS 스마트시키는 웹데스 SOLETHISI © OSS, J ARE_CRAFT(1) © AM, N © 40, AN(1)) © FL STAFT Offset 1541 1543 1473 1479 1429 1429 1420 1426	[활용성용 당보] 봄이나 총 전비: 바반가 동강할 것 PARTI2 ♥ 0.056, PARTI2 ♥ 0.05	다 담은 제조업의 범스, 제품과 등 45 으로 M인다. 물론의 MERAIN(1) 등 06,0 act) 등 ct, 12,44(1) UREADLE(1) 등 ct, 12,44(1) UREADLE(1) 등 ct, 12,44(1) 2014-04-17 2014-04-17 2014-04-17 2014-04-17	27 약이는 가능한 가능한 L은 환이에 도원을 가지 Seleves 이 지, Seleves 이 지, Seleves 이 지, Seleves Annetator admin admin admin admin admin	이 영었다. 20 3.3 있다가라며 Botter net: 44 pro- 44 pro- 44 pro- 44 pro- 44 pro- 44 pro- 44 pro- 55555 844 844 844 844 844 844	대표시계 시장 '장기격으로 tv next ዙ (2) 는 cv_Poucy (5) 는 qt_court HM_COUL_TINUS 는 LINDOW

Figure 5: Annotations results UX

- add/remove a Document to a training/testing set
- register new Annotators Engines
- train/evaluate an Annotator
- test a given Annotator for a given text
- correct a badly tagged Document
- add/remove Resource to Dictionaries

Ultimately, the system generates new entities or facts not explicitly included in the input data, so that Annotator Engines have a better chance of analyzing them correctly, and increasing the coverage of the KB.

## **Experimental Evaluation**

The starting corpus was the Korean version of Wikipedia containing 287,466 entries mapped to the XB Ontology.

By applying source acquisition and expansion, adding several famous Korean online encyclopedias, the Encyclopedia Britannica and nearly 2000 further dictionaries including Wiktionary, the number of documents rose to just under 3,000,000 documents, nearly ten times the size of the original corpus. Like Schlaefer et al. (2011), we cannot at this stage measure the impact of the project on the accuracy and precision of the full QA system, but through the brochettes (sets of sentences, semantically related to an instance of the ontology), human curators were able to triple the speed of discovery of new entities or facts compared to the previous approach of annotating only Wikipedia pages.

It shows that the semantic matching proposes a high precision relation of sentences to given entities, while retaining the discovery of new facts about this entity. Further experiments are still needed to quantify the precision recall of this method. That can be improved using more advanced features and maybe also using a machine learner approach.

## Conclusion

In this paper, we described a platform supporting continuous acquisition and seamless integration of knowledge from various sources and able to integrate various Annotator Engines for extracting new knowledge from unstructured data into a formal Knowledge Base helping to boost quality of the overall QA system.

This is an ongoing work and it is still early to evaluate it in terms of precision / recall or coverage of the full QA system itself, but such a platform definitively boost the productivity of manual tasks for human curators and helps engineers to visualize the performance of their engines. Further work will be done to quantify the quality of suggestion of new knowledge (entities or facts) and coverage of QA system for a given question set.

## Acknowledgements

This work was supported by the Industrial Strategic Technology Development Program (10044494, WiseKB: Big data based self-evolving knowledge base and reasoning platform) funded by the Ministry of Science, ICT & Future Planning (MSIP, Korea)

## References

N. Schlaefer, J. Chu-Carroll, E. Nyberg, J. Fan, W. Zadrozny, and D. Ferrucci. Statistical source expansion for question answering, in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, 2011.

D. Ferrucci and A. Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering, 10(3-4):327–348, 2004.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. : Building Watson: An Overview of the DeepQA Project. AI Mag, Vol 31 (2010) 59–79

F. M. Suchanek, G. Kasneci and G. Weikum. : YAGO: A Core of Semantic Knowledge. Proceedings of the 16th international conference on World Wide Web. ACM New York (2007) 697-706.

Mendes et al., DBpedia spotlight: shedding light on the web of documents, Proceedings of the 7th International Conference on Semantics Systems. ACM(2011).

Aesun Yoon, Soonhee Hwang, Eunroung Lee, and Hyuk-Chul Kwon. : Construction of Korean wordnet KorLex 1.5. Journal of KIISE: Software and Applications, Vol 36(1). (2009) 92–108.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In AAAI.