Linked Data for NLP or by NLP?

Key-Sun Choi
KAIST
kschoi@kaist.ac.kr
http://semanticweb.kaist.ac.kr/

8-OCT-2012 at UKP, TU Darmstadt
Personal History

• NEC C&C Lab. – PIVOT Japanese-Korean Machine Translation
• English-Korean Machine Translation, KAIST+SERI (ETR)
• Korean Part-of-Speech Tagset, Corpus, Dictionary
• KORTERM: Korea Terminology Research Center for Language and Knowledge Engineering (1998-2007)
• CS Department Head, KAIST (2006-2011)
• CoreOnto: ontology for IT area
• Wikipedia to Ontology Cloud, DBPedia Synchronization
• http://lod2.eu/ partner (EU FP7)
Introduction

LINKED OPEN DATA
From the document web to the semantic web

Data Web (since 2006)
- URI de-referencability
- Web Data integration
- RDF serializations

Social Web (since 2003)
- Folksonomies/Tagging
- Reputation, sharing
- Groups, relationships

Semantic Web (Vision 1998, starting ???)
- Reasoning
- Logic, Rules
- Trust

Web (since 1992)
- HTTP
- HTML/CSS/JavaScript

From the slides of LOD2, http://lod2.eu
Collaboratively Generated Contents are...

- **Primary goal**: Knowledge/Information sharing
- Publicly available
- Constructed through explicit/implicit collaboration of multiple authors
  - Subset of User Generated Contents (UGC), such as social nets, emails, etc.

Slide from: http://ir.mathcs.emory.edu/cgc_tutorial_aaai2001
Vast amounts of CGC, in a variety of forms.
Emerging Web of Data

Key: Sun Choi, Linked Data for or by NLP
Wikipedia is **open, collaborative** encyclopedia.

- It is kept updated manually by its collaborators
- Result of collective intelligence: Reflects current trends & culture
- Various links and hypertexts between articles provide additional information
- Multilingual information exists for some contents
- All editing histories are preserved
- New articles are constantly being introduced
## Format of Wikipedia

### Entry/Article title

**Thailand**

From Wikipedia, the free encyclopedia

(Redirected from Siam)

"Siam" redirects here. For other uses, see Siam (disambiguation).

### Redirections

**Thailand** (ภาษาไทย: ประเทศไทย, phonetic: *[nɛ̀ːtlaem]/tʰaːn-lənd or *[nɛ̀ːtlaem]/tʰaːnland;[6] Thai: ประเทศไทย) is a country located at the center of Southeast Asia. It is bordered to the north by Burma and Laos, to the east by Laos and Cambodia, to the south by the Gulf of Thailand and Malaysia, and to the west by the Andaman Sea and the southern extremity of Burma. Its maritime boundaries include Vietnam in the Gulf of Thailand to the southeast and Indonesia and India in the Andaman Sea to the southwest.

The country is a kingdom, with most recorded reigns in the world; a constitutional monarchy with King Rama IX, the ninth king of the House of Chakri, who has reigned since 1946, making him the world’s longest-serving current head of state and the longest-reigning monarch in Thai history.[7] The king is officially titled Head of State, the Head of the Armed Forces, an Upholder of the Buddhist

### Infobox

**Kingdom of Thailand**

Ratcha Anachak Thai
Prathet Thai

Flag

Emblem

Anthem: Phleng Chat Thai

Royal anthem: Phleng Sansoen Phra Barami

### Body

**Categories**: Thailand | Southeast Asian countries | Countries of the Indian Ocean | Countries bordering the South China Sea | Constitutional monarchies | Member states of the Association of Southeast Asian Nations | Member states of the United Nations

---

[1] Sun Choi, Linked Data for or by NLP
DBpedia

- DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web (Definition From: http://wiki.dbpedia.org/About)

- Elements of DBpedia:
  - RDF triples, which are direct conversion of Wikipedia Infoboxes (Core Element)
  - Category information and short abstracts retrieved from Wikipedia
  - Manual links to external Wikipedia-based ontologies, such as YAGO
Paradigm of Linked Data

- Uses RDF as data model

  - Korean Government funds KAIST
  - KAIST hasPresident Nam-Pyo Suh
  - KAIST locatedAt Daejeon

- Is serialized in triples

  - Korean Government funds KAIST
  - KAIST hasPresident Nam-Pyo Suh
  - KAIST locatedAt Daejeon
## Kingdom of Thailand

<table>
<thead>
<tr>
<th>Capital (and largest city)</th>
<th>Bangkok¹ (in Thai: Krung Thep)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official language(s)</td>
<td>Thai</td>
</tr>
<tr>
<td>Official scripts</td>
<td>Thai alphabet</td>
</tr>
<tr>
<td>Demonym</td>
<td>Thai</td>
</tr>
<tr>
<td>Government</td>
<td>Unitary parliamentary democracy and Constitutional monarchy</td>
</tr>
<tr>
<td>- King</td>
<td>Bhumibol Adulyadej</td>
</tr>
<tr>
<td>- Prime Minister</td>
<td>Yingluck Shinawatra</td>
</tr>
</tbody>
</table>

### Area

- Total: 513,120 km² (51st)
- Water (%): 0.4 (2,230 km²)
Construction of DBpedia

<table>
<thead>
<tr>
<th>dbpedia: Thailand</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:capital</td>
</tr>
<tr>
<td>dbpedia-owl:language</td>
</tr>
<tr>
<td>dbpedia:Thai_alphabet</td>
</tr>
<tr>
<td>dbpedia-owl:denonym</td>
</tr>
<tr>
<td>dbpedia-owl:governmentType</td>
</tr>
<tr>
<td>dbpedia:Parliamentary_system</td>
</tr>
<tr>
<td>dbpedia-owl:leaderName</td>
</tr>
<tr>
<td>dbpedia:Parliamentary_system</td>
</tr>
<tr>
<td>dbpprop:areaKm</td>
</tr>
<tr>
<td>dbpprop:areaSqMi</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Capital (and largest city)</th>
<th>Bangkok³ (in Thai: Krung Thep)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official language(s)</td>
<td>Thai²¹</td>
</tr>
<tr>
<td>Official scripts</td>
<td>Thai alphabet</td>
</tr>
<tr>
<td>Demonym</td>
<td>Thai</td>
</tr>
<tr>
<td>Government</td>
<td>Unitary parliamentary democracy and Constitutional monarchy</td>
</tr>
<tr>
<td>- King</td>
<td>Bhumibol Adulyadej</td>
</tr>
<tr>
<td>- Prime Minister</td>
<td>Yingluck Shinawatra</td>
</tr>
<tr>
<td>Area</td>
<td>513,120 km² (51st) 198,115 sq mi</td>
</tr>
<tr>
<td>- Total</td>
<td>513,120 km²</td>
</tr>
<tr>
<td>- Water (%)</td>
<td>0.4 (2,230 km²)</td>
</tr>
</tbody>
</table>
The web of documents

- The Web is a single information space build on open standards and hyperlinks.

Linked Data

- Use RDF and HTTP to
  1. Publish structured data on the Web,
  2. Set data links between data from one data source to data within other data sources.

Getting Involved to LOD

- As a data publisher, you need to...
  - Convert some data to RDF
  - Publish it on your website like you do with HTML

- As a data consumer, you need to...
  - Learn one of the open source or commercial RDF toolkits
  - Find and retrieve the data you need
  - Use SPARQL to extract the useful bits for the task at hand

Slide from: Ian Davis and Tom Heath, The thirty minute guide to RDF and Linked Data from slideshare.
Applications of DBpedia

- Improving Wikipedia search
  - DBpedia supports SPARQL query endpoints for structured part of Wikipedia
  - Enable sophisticated queries to Wikipedia, such as:
    - Rivers that flow into the Rhine and are longer than 50 kilometers
    - Albums from the Beach Boys that were released between 1980 and 1990
    - French scientists who were born in the 19th century
    - Skyscrapers in China that have more than 50 floors
    - Actors of the American TV-series Lost that were born before 1970
LOD by NLP

NLP2RDF

Link with DBPedia

Multilingual Synchronization focusing on DBpedia/Wikipedia
Table

1. Project progress
2. NLP2RDF
3. Link with DBpedia
1. Project progress

2. NLP2RDF – WP3

3. Link with DBpedia – WP3

4. Semi-Automatic Data Interlinking – WP4

5. IASLOD 2012 – WP10
Scope of the Project

**LOD2 Work Packages**

**Annex I. Description of Work**

**WP1: Requirements, design and LOD2 stack prototype**
- WP2: Storing and Querying Large Knowledge bases
- WP3: Enrichment, Classification and Interlinking
- WP4: Knowledge Fusion and Information Quality
- WP5: Visualisation, Browsing & Authoring

**WP6: Interfaces, component integration and LOD2 stack**
- WP7: LOD2 for Media & Publishing
- WP8: LOD2 4 Enterprise Data Web
- WP9: LOD2 for Government Data
- WP9a: LOD2 for Public Contracts

**WP10: Training, dissemination, community building, fertilisation**

**WP11: Exploitation & standards**

**WP12: Project Management**

**Figure 4: Interrelation and information flow between individual LOD2 work packages.**
LOD2 Stack Architecture
Project progress (2012.Sep, KAIST)

**NLP2RDF**
- NLP tool
- Wrapper
- Ontology
- NIF

**Link Discovery**
- DBpedia
- Semi-Automatic Link
- Automatic Link

**Link to LOD**
- LOD2 Stack

**English**
- Stanford Core-NLP
- Stanford Core-NLP Wrapper
- DBpedia Spotlight
- SILK
- LIMES
- Early Stack Prototype

**Korean**
- Korean NLP2RDF Platform
- String Ontology
- OLiA
- Korean DBpedia Framework
- Korean Localized SILK
- Link with wiki-article
1. Project progress

2. NLP2RDF – WP3

3. Link with DBpedia – WP3

4. Semi-Automatic Data Interlinking – WP4

5. IASLOD 2012 – WP10
NLP2RDF

- LOD2 Community project
- Developing the **NLP Interchange Format (NIF)**
- NIF aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations
NLP2RDF - English

- **NIF**
  - NIF 1.0 (in Sept. 2011)
  - String Ontology
  - NIF 2.0 Draft

- **Implementations**
  - Gate ANNIE
  - DBpedia Spotlight
  - MontyLingua
  - Stanford CoreNLP
  - SnowballStemmer
  - FOX
  - …
  - And, Wrappers

- **Flow**

  1. Raw Text
  2. NLP tool
  3. Output
  4. Wrapper
  5. NIF output
     *RDF/OWL*
NLP2RDF – Korean

• Implementations

  • HanNanum
    ✓ Open Source Korean Morpheme Analyzer
    ✓ Developed by SWRC

  • Korean Berkeley Parser *
    ✓ F1-score: 77.29%

  • Wrapper
    ✓ Compliant with the NIF 1.0
    ✓ Korean Tag Set (Sejong Tag Set) Linked with OLiA
    ✓ Produce RDF triples

• Korean NLP2RDF Platform

  Raw Text → NLP tools → Morpheme Analyzer → Parser → Output → Wrapper → NIF output

* RDF/OWL
NLP2RDF – Korean
Korean Parser (1/2)

- **Korean parser**
  - Aim to build a Korean syntactic parser.
  - For the first step, try to train the existing English parsers to run on Korean texts

- **Refining Corpus**
  - Using Sejong Treebank
  - Convert Sejong Treebank into Penn-Treebank format

- **Training to existing three PSG parser**
  - Stanford parser // F1-score: 73.92%
  - Berkeley parser // F1-score: 72.35%
  - Bikel-Collins parser // F1-score: 77.29%

- Berkeley parser worked best for Korean
- In NLP2RDF and LOD work, Berkeley parser is used

* DongHyun Choi, Jungyeul Park, Key-Sun Choi, Korean Treebank Transformation for Parsr Training, ACL - SPMRL 2012
NLP2RDF – Korean Parser (2/2)

- **Sejong Treebank**
  - Eojeol-based (use eojeols as its fundamental unit of analysis)
  - Each eojeol could contain one or more morphemes which have very different grammatical roles
  - In most case, eojeols are determined by white spaces

- **Transforming Methods from Eojeol-based to Entity-based**
  For example:

<table>
<thead>
<tr>
<th>System</th>
<th>Corpus</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>67.88%</td>
<td>61.77%</td>
<td>64.69%</td>
<td></td>
</tr>
<tr>
<td>M 1</td>
<td>68.34%</td>
<td>61.93%</td>
<td>64.98%</td>
<td></td>
</tr>
<tr>
<td>M 1-2</td>
<td>71.78%</td>
<td>67.50%</td>
<td>69.58%</td>
<td></td>
</tr>
<tr>
<td>M 1-3</td>
<td>71.28%</td>
<td>67.91%</td>
<td>69.56%</td>
<td></td>
</tr>
<tr>
<td>M 1-4</td>
<td>71.06%</td>
<td>67.08%</td>
<td>69.01%</td>
<td></td>
</tr>
<tr>
<td>M 1-5</td>
<td>71.35%</td>
<td>67.27%</td>
<td>69.26%</td>
<td></td>
</tr>
<tr>
<td>M 1-6</td>
<td>75.85%</td>
<td>72.07%</td>
<td>73.92%</td>
<td></td>
</tr>
</tbody>
</table>

![Diagram of transforming methods from Eojeol-based to Entity-based]
NLP2RDF – Korean to Compliant with the NIF

• Use Hash-based URI

• Use NIF Vocabulary

• Use OLiA

“Korean Linked Data on the Web: From Text to RDF,” JIST (Joint International Semantic Technology Conference), 2012, Nara - Martín Rezk, Jungyeul Park, Yoon Yongun, Kyungtae Lim, John Larsen, YoungGyun Hahm, and Key-Sun Choi
NLP2RDF – Korean
- Hash-Based URI

- **Example Sentence** *(Morpheme Analyzing and POS tagging)*
- “KAIST는 LOD2 프로젝트에 참여하고 있다.”
  - “KAIST is participating in the LOD2 project.”

<table>
<thead>
<tr>
<th>Text</th>
<th>POS</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAIST</td>
<td>SL</td>
<td>hash_15_1.66666666666667_44ab5b1e80735b9bf793bb03f4b9b81_KAIST</td>
</tr>
<tr>
<td>는</td>
<td>JK</td>
<td>hash_15_1.d9747f9c4821ffaa4e5e6af6d49e06.는</td>
</tr>
<tr>
<td>LOD2</td>
<td>SL</td>
<td>hash_15_1.3333333333333_7ca8eb8f8ce4475fa6c4307e52ce09e LOD2</td>
</tr>
<tr>
<td>프로젝트에</td>
<td>NNG</td>
<td>hash_15_4.ae2f21dfb7b9cd223f3d82cd480675 프로젝트</td>
</tr>
<tr>
<td>이</td>
<td>JKB</td>
<td>hash_15_1.31cd4e7e385e1aaf90185d67b3c6567a_에</td>
</tr>
<tr>
<td>참여</td>
<td>NNG</td>
<td>hash_15_2.bd99953d9e1c37b7ca4ed96ac75363_참여</td>
</tr>
<tr>
<td>하고</td>
<td>JKB</td>
<td>hash_15_2.6ce14645d0c30e26824504b9071ca6d6 하고</td>
</tr>
<tr>
<td>있</td>
<td>VX</td>
<td>hash_15_1.72810cd5156c302e9e8fc3e0798aad4_있</td>
</tr>
<tr>
<td>다</td>
<td>EF</td>
<td>hash_15_1.0dd7e4390635fb75323ad5fc353ea2 다</td>
</tr>
<tr>
<td>..</td>
<td>SF</td>
<td>hash_15_1.079c140f9653ee8235753b72d100371a..</td>
</tr>
</tbody>
</table>

KAIST는 LOD2 프로젝트에 참여하고 있다.  ROOT  hash_0.20_3c30e32e75b90d64465f23e721435fKAIST는%20LOD2%20프로젝트
KAIST는 LOD2 프로젝트에 참여하고 있다.  S  hash_0.25.883962a3bbe0dabc80a7b4e45b8adb81KAIST는%20LOD2%20프로젝트
KAIST는 LOD2 프로젝트에 참여하고 있다.  NP  hash_0.3.68666666666667_001cb1ae2e571c96ee3b4522023a77KAIST는%20
 LOD2 프로젝트에 참여하고 있다.  VP  hash_0.20.3333333333333_ecl19f8b9703a9170f53b4d1f73b14_LOD2%20프로젝트에%20
 LOD2 프로젝트에  NP-AJT  hash_0.9.3333333333333_e00035e52e13715b55267dfe06d3d1_LOD2%20프로젝트에%20
 LOD2          | NP  | hash_0.2.3333333333333_e55903b707c0aee9ae56e6d1cf496aa7 LOD2%20 |
| 프로젝트에    | NP  | hash_0.6_ff4c6d384888dccc62210569387b1 프로젝트에%20 |
| 참여하고 있다. | VP  | hash_0.10.59a0c169aa9de8dd9959678bb28bad4 참여하고%20 |
| 참여하고     | NP  | hash_0.5_5ce157229be1ddc09d85608a8d865 참여하고%20 |
| 있다.        | VP  | hash_0.4_43e19a7866428787c7730c83b623758. 있다. |

Key-Sun Choi, Linked Data for or by NLP
NLP2RDF – Use NIF vocabulary

- http://nlp2rdf.lod2.eu/schema/string/
  - ContextHashBasedString
  - Document
  - OffsetBasedString
  - Phrase
  - Sentence
  - Word

- Object Property
  - nextSentenceTrans
  - nextWordTrans
  - olaLink
  - previousSentenceTrans
  - previousWordTrans
  - sourceUrl
  - substringTrans
  - superStringTrans

- Data Property
  - anchorOf
  - beginIndex
  - endIndex
  - leftContext
  - lemma
  - posTag
  - rightContext
  - stem

*temporarily added this declaration.*
## NLP2RDF – Korean
### Link Sejong Tagset with OLiA (1/3)

- **English**
  - Penn-OLiA Link

- **Korean**
  - Sejong-OLiA Link

<table>
<thead>
<tr>
<th>채언</th>
<th>영문</th>
<th>한문</th>
<th>유언</th>
</tr>
</thead>
<tbody>
<tr>
<td>명사</td>
<td>NN</td>
<td>일반명사</td>
<td>경험명사</td>
</tr>
<tr>
<td>대명사</td>
<td>NP</td>
<td>고유명사</td>
<td>NNP</td>
</tr>
<tr>
<td>수사</td>
<td>NR</td>
<td>이존명사</td>
<td>NNB</td>
</tr>
<tr>
<td>동사</td>
<td>VV</td>
<td>대명사</td>
<td>NNP</td>
</tr>
<tr>
<td>명사</td>
<td>VA</td>
<td>대명사</td>
<td>NNP</td>
</tr>
<tr>
<td>보조용언</td>
<td>VX</td>
<td>경험명사</td>
<td>VCP</td>
</tr>
<tr>
<td>지정사</td>
<td>VC</td>
<td>부정지정사</td>
<td>VCN</td>
</tr>
<tr>
<td>관형사</td>
<td>MM</td>
<td>경험사</td>
<td>MM</td>
</tr>
<tr>
<td>부사</td>
<td>MA</td>
<td>일반부사</td>
<td>MAG</td>
</tr>
<tr>
<td>명사</td>
<td>IC</td>
<td>접속부사</td>
<td>MAJ</td>
</tr>
<tr>
<td>관계언</td>
<td>JK</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>격조사</td>
<td>JX</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>관형언</td>
<td>E</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>어미</td>
<td>EP</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>의존형태</td>
<td>XP</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>첨두사</td>
<td>E</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>접미사</td>
<td>XS</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>어근</td>
<td>XR</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>기호</td>
<td>SF</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>심태</td>
<td>SP</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>다형</td>
<td>SS</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>볼일</td>
<td>SE</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>볼임</td>
<td>SO</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>볼자</td>
<td>SH</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>기타기호</td>
<td>SW</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>복수</td>
<td>SF</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>복수형</td>
<td>SP</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>복수형</td>
<td>SS</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>복수형</td>
<td>SE</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>복수형</td>
<td>SO</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>복수형</td>
<td>SH</td>
<td>접속조사</td>
<td>IC</td>
</tr>
<tr>
<td>복수형</td>
<td>SW</td>
<td>접속조사</td>
<td>IC</td>
</tr>
</tbody>
</table>

Key-Sun Choi, Linked Data for or by NLP
### NLP2RDF – Korea

: Link Sejong Tagset with OLiA (2/3)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Sejong</th>
<th>OLiA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LinguisticAnnotation/Tag/</td>
<td>LinguisticConcept/MorphosyntacticCategory/</td>
</tr>
<tr>
<td>MA</td>
<td>Adverb</td>
<td>Adverb</td>
</tr>
<tr>
<td>MAJ</td>
<td>Adverb/ConjunctiveAdverb</td>
<td>Adverb and Conjunction/CoordinatingConjunction</td>
</tr>
<tr>
<td>MAG</td>
<td>Adverb/GeneralAdverb</td>
<td>Adverb</td>
</tr>
<tr>
<td>SN, XN</td>
<td>CardinalNumber</td>
<td>Quantifier/Numeral</td>
</tr>
<tr>
<td>MM</td>
<td>Determiner</td>
<td>PronounOrDeterminer/Determiner</td>
</tr>
<tr>
<td>SH, SL</td>
<td>ForeignWord</td>
<td>Residual/Foreign</td>
</tr>
<tr>
<td>IC</td>
<td>Interjection</td>
<td>Interjection</td>
</tr>
<tr>
<td>XR</td>
<td>Noun/BaseMorpheme</td>
<td>Noun/CommonNoun</td>
</tr>
<tr>
<td>NN</td>
<td>Noun</td>
<td>Noun</td>
</tr>
<tr>
<td>NNB, NNG</td>
<td>Noun/CommonNoun</td>
<td>Noun/CommonNoun</td>
</tr>
<tr>
<td>NNP</td>
<td>Noun/ProperNoun</td>
<td>Noun/ProperNoun</td>
</tr>
<tr>
<td>NA, NF</td>
<td>LikelyNoun</td>
<td>Noun</td>
</tr>
<tr>
<td>NP</td>
<td>Pronoun</td>
<td>PronounOrDeterminer/Pronoun</td>
</tr>
<tr>
<td>SE, SF, SO, SP, SS</td>
<td>Symbol</td>
<td>Punctuation</td>
</tr>
<tr>
<td>NV, V</td>
<td>Verb</td>
<td>Verb</td>
</tr>
<tr>
<td>VA</td>
<td>Verb/Adjective</td>
<td>Verb and Adjective/PredicativeAdjective</td>
</tr>
<tr>
<td>VX</td>
<td>Verb/AuxiliaryPredicate</td>
<td>Verb/AuxiliaryVerb</td>
</tr>
<tr>
<td>VC, VCN, VCP</td>
<td>Verb/Copula</td>
<td>Verb/FiniteVerb</td>
</tr>
<tr>
<td>VV</td>
<td>Verb/VerbalPredicate</td>
<td>Verb</td>
</tr>
<tr>
<td>E, JK, XP, XS</td>
<td>Particle</td>
<td>MorphologicalCategory/morpheme/</td>
</tr>
<tr>
<td>JC, JX</td>
<td>Particle/AuxiliaryPostposition</td>
<td>MorphologicalCategory/morpheme/Morphological</td>
</tr>
<tr>
<td>JKB, JKC, JKG, JKO, JKQ, JKS, JKV</td>
<td>Particle/CaseMarker</td>
<td>MorphologicalCategory/morpheme/Morphological</td>
</tr>
<tr>
<td>XPN</td>
<td>Particle/Prefix</td>
<td>MorphologicalCategory/morpheme/Prefix</td>
</tr>
<tr>
<td>XSA, XSN, XSV</td>
<td>Particle/Suffix</td>
<td>MorphologicalCategory/morpheme/Suffix</td>
</tr>
<tr>
<td>EC, EF, EP, ETM, ETN</td>
<td>Particle/VerbalEnding</td>
<td>MorphologicalCategory/morpheme/verbalEnding</td>
</tr>
</tbody>
</table>
NLP2RDF – Korean
Link Sejong Tagset with OLiA (3/3)

• Discrepancy between Penn and Sejong tagset with OLiA
  
  • Change original Sejong Tagset structure to likely Penn Tagset structure
    • For multilingual interchange
  • Korean specific grammatical feature
    • Josa (Postposition – Case Marker)
      • MorphologicalCategory/Morpheme/
    • Adjective
      • Used like Verb in Korean
      • So linked with Verb and Adjective/PredicativeAdjective
    • BaseMorpheme
      • In original Sejong Tagset: Particle
      • Here: CommonNoun (actually it is considered ‘word’)
NLP2RDF – Korean Demo

- [http://semanticweb.kaist.ac.kr/nlp2rdf/](http://semanticweb.kaist.ac.kr/nlp2rdf/)
- “KAIST는 LOD2 프로젝트에 참여하고 있다.”
Korean NLP2RDF

**RAW Text**
- **HanNanum**
  - Korean Open Source Morpheme Analyzer
  - Developed by SWRC, KAIST

**Morphological Analyzer**
- **Korean Berkeley Parser**
  - Training set: Modified Sejong Treebank (DongHyun Choi, Jungyeul Park, Key-Sun Choi, *Korean Treebank Transformation for Parsr Training*, ACL - SPMRL 2012)
  - F1-score: 82.12%

**Parser**

**Wrapper**
- **Produce triples**
  - Use OLiA (Ontologies of Linguistic Annotation) to link the Korean tagsets with language-independent reference concepts
  - The OLiA annotation model and the OLiA linking model produce triples using the Sejong tagset
1. Project progress

2. NLP2RDF – WP3

3. Link with DBpedia – WP3

4. Semi-Automatic Data Interlinking – WP4
**Link with DBpedia – WP3**

- **DBpedia** is a community effort to extract structured information from *Wikipedia* and to make this information available on the Web.
- DBpedia allows you to ask sophisticated queries against Wikipedia, and to link other data sets on the Web to Wikipedia data.

  - Build Korean DBpedia
  - SPARQL endpoint

---

Korean DBpedia dataset

<table>
<thead>
<tr>
<th>URL-encoding</th>
<th>Extractor</th>
<th>nt-file</th>
<th>csv-file</th>
</tr>
</thead>
<tbody>
<tr>
<td>KoreanDBpedia</td>
<td>ArticleTitle</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>ArticleCategories</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>ArticleLabel</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>Category</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>Description</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>ExternalLink</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>GeoCoordinates</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>HomePage</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>Image</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>InfoBox</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>InfoBoxLabel</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>Page</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>PageTitle</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>SKOSCategories</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>KoreanDBpedia</td>
<td>Wikibase</td>
<td>nt @</td>
<td>csv @</td>
</tr>
</tbody>
</table>

---

IRI-encoding

<table>
<thead>
<tr>
<th>Extractor</th>
<th>nt-file</th>
<th>csv-file</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArticleTitle</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>ArticleCategories</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>ArticleLabel</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>Category</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>Description</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>ExternalLink</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>GeoCoordinates</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>HomePage</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>Image</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>InfoBox</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>InfoBoxLabel</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>Page</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>PageTitle</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>SKOSCategories</td>
<td>nt @</td>
<td>csv @</td>
</tr>
<tr>
<td>Wikibase</td>
<td>nt @</td>
<td>csv @</td>
</tr>
</tbody>
</table>
Link with DBpedia
Wikipedia sentence and infobox triple mapping in DB (1/2)

- Infobox is based on Wikipedia article

- DBpedia triple can be considered
  ‘Subject’ ‘Predicate’ ‘Object’ of Wikipedia sentence

- To Link with DBpedia,
  - Our first step: Make DB- mapping between Wikipedia and infobox triple
Link with DBpedia - Korean Wikipedia sentence and infobox triple mapping in DB (2/2)

<table>
<thead>
<tr>
<th>doc_title</th>
<th>sent_text</th>
<th>subject</th>
<th>predicate</th>
<th>object</th>
</tr>
</thead>
<tbody>
<tr>
<td>김효일</td>
<td>김효일(1979년 9월 7일 ~)은 대한민국의 축구 선수로서 포지션은 미드필더이다.</td>
<td>김효일</td>
<td>포지션</td>
<td>미드필더</td>
</tr>
<tr>
<td>제마우. 공항</td>
<td>제마우 공항 ( IATA: CAH, ICAO: YVON )은 베트남 젤롱월이 성 광용이에 위치한 공항</td>
<td>제마우 공항</td>
<td>ICAO: YVON</td>
<td>베트남 젤隆월이성광용이에 위치한공항</td>
</tr>
<tr>
<td>제마우. 공항</td>
<td>제마우 공항 ( IATA: CAH, ICAO: YVON )은 베트남 젤隆월이 성 광용이에 위치한 공항</td>
<td>제마우 공항</td>
<td>ICAO: YVON</td>
<td>베트남 젤隆月이성광용이에 위치한공항</td>
</tr>
<tr>
<td>혜명예문</td>
<td>혜명예문은 마자라우고과속하는 마라헤말과로 종합은 Jeffersonia dubia 이다.</td>
<td>혜명예문</td>
<td>ICAO: YVON</td>
<td>제단속해진해.data</td>
</tr>
<tr>
<td>컨터. 공항</td>
<td>컨터 공항 ( IATA: VCA, ICAO: VCTR )은 베트남 젤隆월이성광용이에 있는 공항</td>
<td>컨터. 공항</td>
<td>ICAO: VCTR</td>
<td>베트남 젤隆月이성광용이에 있는공항</td>
</tr>
<tr>
<td>클러치</td>
<td>클러치는 종종미에 속하여 학명은 Lepisota tanaka이다.</td>
<td>클러치</td>
<td>학명</td>
<td>Lepisota tanaka</td>
</tr>
<tr>
<td>클러치</td>
<td>클러치는 종종미에 속하여 학명은 Cololabis saira이다.</td>
<td>클러치</td>
<td>학명</td>
<td>Cololabis saira</td>
</tr>
<tr>
<td>클러치</td>
<td>클러치는 종종미에 속하여 학명은 Cololabis saira이다.</td>
<td>클러치</td>
<td>학명</td>
<td>Cololabis saira</td>
</tr>
<tr>
<td>클러치</td>
<td>클러치는 종종미에 속하여 학명은 Cololabis saira이다.</td>
<td>클러치</td>
<td>학명</td>
<td>Cololabis saira</td>
</tr>
<tr>
<td>클러치</td>
<td>클러치는 종종미에 속하여 학명은 Cololabis saira이다.</td>
<td>클러치</td>
<td>학명</td>
<td>Cololabis saira</td>
</tr>
</tbody>
</table>

주석

- Cololabis saira: 미국 통행문류정보시스템(ITIS 19 March 2008)에 확인.
Link with DBpedia for Example- DBpedia spotlight

- Annotates text with links to LOD via DBpedia
- Recognizes entities (Person, Location, Organization)
- Also links entity mentions to unique identifiers

President Obama on Monday will call for a new minimum tax rate for individuals making more than $1 million a year ensuring that they pay at least the same percentage of their earnings as other taxpayers, according to administration officials.
행정부 관리의 공식 발언을 따르면, 월요일에 오바마 대통령은 1년에 1백만 달러 이상을 확보하기 위하여 최소한 다른 납세자들처럼 수익의 동일한 비율만큼을 지불하도록 하는 새로운 최소 세율을 요청할 것이다.

"President Obama on Monday will call for a new minimum tax rate for individuals making more than $1 million a year to ensure that they pay at least the same percentage of their earnings as other taxpayers, according to administration officials.”
Link with DBpedia to Annotate text with links to LOD (2/3)

- Now,
  - Find exact matching string
  - Link with Wikipedia page
  - Experiment on Scientific articles
    - Easy to find in Wikipedia
    - Relatively Less ambiguity

- Ambiguity
  - Synonym
  - Homonym

http://semanticweb.kaist.ac.kr/nlp2rdf/
http://ko.dbpedia.org/sparql/
Link with DBpedia - Korea to Annotate text with links to LOD (3/3)

- Remaining…
  - Disambiguation
    - SILK
    - LIMES
  - Generality
    - Articles in a variety of genres
  - Link to LOD via DBpedia
    - en.dbpedia.org / ko.dbpedia.org
  - Semi-Automatic Data Interlinking
    - SILK
1. Project progress

2. NLP2RDF – WP3

3. Link with DBpedia – WP3

4. Semi-Automatic Data Interlinking – WP4

5. IASLOD 2012 – WP10
Semi-Automatic Data Interlinking

- Silk
  - Silk GUI Localization
  - String Similarity Measure Localization
    - Similarity measure between Korean strings
    - Similarity measure between transliterated Korean strings

- Link to LOD via DBpedia
  - Test data set: Korean recipe data & DBpedia
Semi-Automatic Data Interlinking

SILK - Overall Architecture

Source Data in relational

RDF generator

Instances

Ontology

Overall Architecture

RDF Links

Linked Korean Recipes

DBpedia

Key-Sun Choi, Linked Data for or by NLP
Semi-Automatic Data Interlinking
SILK GUI Korean Localization

- Decode encoded URIs in SILK
  - For display
  - For human user

- Displaying and Storing Double-Byte Characters
The scale of comparison is different
- Latin or Roman alphabets / Korean alphabets

We implemented `KoreanPhonemeDistance` operator
- Based on the level of distribution of Korean phonemes

KPD (Korean Phoneme Distance) operator
- Existing Korean algorithms (kored)
  - *Does not consider syllables*
- KPD
  - *Weight different syllables*

For example
- English: `ramble` / `tumbler`
- Korean: `한국` / `함구령`
Semi-Automatic Data Interlinking: Similarity Measures for Double-Byte Characters (2/2)

Application of Traditional Edit Distance to Korean Resources

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Levenshtein Distance</th>
<th>Actual Edit Operation</th>
<th>Differences in phonemes</th>
<th>Differences in syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>녹차</td>
<td>모과차</td>
<td>2</td>
<td>3 (ㅁ→ㄴ, ㄱ→ add, 과→ delete)</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Comparison of Similarity Measures for Korean

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Levenshtein Distance</th>
<th>KorED</th>
<th>GrpSim</th>
<th>OneDSim2</th>
<th>KorPhoD</th>
</tr>
</thead>
<tbody>
<tr>
<td>우연히</td>
<td>망연이</td>
<td>2</td>
<td>β + α</td>
<td>β + α<strong>w3</strong></td>
<td>β + α</td>
<td></td>
</tr>
<tr>
<td>강낭콩</td>
<td>봉낭콩</td>
<td>2</td>
<td>β + α</td>
<td>β + α<strong>w4</strong></td>
<td>β + α</td>
<td></td>
</tr>
<tr>
<td>일반통계학</td>
<td>일반통행</td>
<td>3</td>
<td>β + 3α</td>
<td>β + α<strong>w+d+α*w+α</strong>w</td>
<td>2β + α</td>
<td></td>
</tr>
<tr>
<td>바람</td>
<td>보름</td>
<td>2</td>
<td>2α</td>
<td></td>
<td></td>
<td>β + α</td>
</tr>
</tbody>
</table>

β: syllable distance, α: phoneme distance

Performance Comparison for DBpedia data

<table>
<thead>
<tr>
<th></th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic</td>
<td>81.80</td>
<td>97.89</td>
<td>88.13</td>
</tr>
<tr>
<td>Levenshtein</td>
<td>21.59</td>
<td>100.00</td>
<td>35.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Retrieved</th>
<th>Ret. &amp; Rel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic</td>
<td>6,852</td>
<td>8,211</td>
<td>6,717</td>
</tr>
<tr>
<td>Levenshtein</td>
<td>31,787</td>
<td>31,862</td>
<td></td>
</tr>
</tbody>
</table>

Threshold = 1

• Precision is improved about our times and F-score is improved two-and-a-half times.
Semi-Automatic Data Interlinking
: Similarity Measures for Transliterated Resources (1/2)

- We implemented *KoreanTranslitDistance* and *KoreanRRdistance* operator

- Soundex algorithm – English
  - Remove the vowels, and consonants is replaced
  - Smith / Smythe $\rightarrow$ S1_23 / S1_23_ $\rightarrow$ Same!

- *KoreanTranslitDistance* - Korean
  - 칠국수 $\rightarrow$ kalguksoo / kalgugsoo
    $\rightarrow$ kalguksoo / kalguksoo $\rightarrow$ Same!

- Notation standards are not kept well
Semi-Automatic Data Interlinking: Similarity Measures for Transliterated Resources (2/2)

## Korean Resources in several writing systems

<table>
<thead>
<tr>
<th>Transliteration Type</th>
<th>Examples</th>
<th>Resource location</th>
</tr>
</thead>
</table>

## Transliteration Algorithms for Korean

<table>
<thead>
<tr>
<th>Transliteration Algorithms</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCune-Reischauer(1937)</td>
<td>Official standard in the past (from 1984 to 2000)</td>
</tr>
</tbody>
</table>

## Performance Comparison for DBpedia data

<table>
<thead>
<tr>
<th>Revised Roman</th>
<th>Relevant</th>
<th>Retrieved</th>
<th>Ret. &amp; Rel.</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levenshtein²</td>
<td>5,342</td>
<td>5,341</td>
<td>99.99</td>
<td>79.90</td>
<td></td>
</tr>
<tr>
<td>Levenshtein¹</td>
<td>7,358</td>
<td>6,176</td>
<td>83.94</td>
<td>92.37</td>
<td></td>
</tr>
<tr>
<td>KoRRdist³</td>
<td>5,742</td>
<td>5,734</td>
<td>99.88</td>
<td>85.76</td>
<td></td>
</tr>
<tr>
<td>KoRRdist¹</td>
<td>7,829</td>
<td>6,359</td>
<td>81.22</td>
<td>95.11</td>
<td></td>
</tr>
</tbody>
</table>

- Precision is similar between Levenshtein and KoRRdist, but recall is improved regardless of threshold.

<table>
<thead>
<tr>
<th>McCune-Reischauer</th>
<th>Relevant</th>
<th>Retrieved</th>
<th>Ret. &amp; Rel.</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levenshtein²</td>
<td>3,037</td>
<td>3,036</td>
<td>99.97</td>
<td>86.21</td>
<td></td>
</tr>
<tr>
<td>Levenshtein¹</td>
<td>5,480</td>
<td>4,726</td>
<td>86.56</td>
<td>71.93</td>
<td></td>
</tr>
<tr>
<td>KoMRdist³</td>
<td>4,380</td>
<td>4,376</td>
<td>99.91</td>
<td>86.81</td>
<td></td>
</tr>
<tr>
<td>KoMRdist¹</td>
<td>6,253</td>
<td>5,315</td>
<td>85.00</td>
<td>80.90</td>
<td></td>
</tr>
</tbody>
</table>

- Precision is similar between Levenshtein and KoMRdist, but recall is improved regardless of threshold.
Semi-Automatic Data Interlinking:
Achievements & Conclusion in SILK

- Now, just
  - SILK has been localized to handle Korean resources
  - Semi-Automatic
  - Experiment on food
    - First step to disambiguation (for Synonym, Homonym)

- Remaining...
  - Chinese & Japanese Measure...
  - Semantic Level Matching
  - Full automation among Different Languages
1. Project progress

2. NLP2RDF – WP3

3. Link with DBpedia – WP3

4. Semi-Automatic Data Interlinking – WP4

5. IASLOD 2012 – WP10
2012 International Asian Summer School on Linked Data (1/3)

- http://semanticweb.kaist.ac.kr/2012lodsummer/
2012 International Asian Summer School on Linked Data (2/3)

Sören Auer (U Leipzig)
Introduction to Linked Data and its Life-Cycle

Key-Sun Choi (KAIST)
LOD2 Korea

Dominic Difranzo (RPI, USA)
Open Government Data with LOD and Semantic Web

Kouji Kozaki (Osaka U)
1) LOD Challenge and Ontology Engineering to Enrich Linked Data
2) Current Trends of Japanese LOD -through LOD Challenge Japan

Pablo Mendes (Free U Berlin)
Building A Web of Linked Entities

Axel Ngonga (U Leipzig)
From Unstructured information to Linked Data

Jeff Z. Pan (U Aberdeen)
1) Stream LOD and Querying Linked Ontological Data
2) Linked Data Enabled Software Engineering

Virach Sornlertlamvanich (NECTEC)
Creating Linked Data from Cultural Knowledge Society

Hideaki Takeda (NII)
1) General Introduction for Semantic Web, Linked Open Data
2) Identity and schema for Linked Open Data
3) LOD Application Exemplar

Fumihiro Kato (NII)
Programming with LOD

Jin-Dong Kim (DBCLS)
Developing and publishing bio-medical data as LOD

In-Young Ko (KAIST)
LOD Localization and Application

MunYong Yi (KAIST)
Interlinking Korean Resources on the Web: Issues in Localization

Young-il Kwon (NIA)
Nation DB Project and Open Government Data in Korea

Sung-guk Hahn (Wonkwang U)

Seungwoo Lee (KISTI)
Development of Medical Term Search System Using Bioinformatics Linked Data

Tony Lee (Saltlux)
Geo-Semantic Applications with LOD
2012 International Asian Summer School on Linked Data (3/3)

Program Committee

Sören Auer (U Leipzig)
Key-Sun Choi (KAIST)
Jin-Dong Kim (DBCLS)
Organizing Committee

Advisory Committee

Seungkoo Hwang (ETRI)
Young-il Kwon (NIA)
Younghhee Lee (KAIST)

Organizing Committee

Key-Sun Choi (KAIST) (Chair)
Mun Y. Yi (KAIST) (Co-Chair)
Sung-Kook Han (Wonkwang U)
Dosam Hwang (Yeungnam U)
Hanmin Jung (KISTI)
Jason J. Jung (Yeungnam U)
In-Young Ko (KAIST)
Uchin Lee (KAIST)
Yong-Un Yoon (KAIST)
Martin Rezk (KAIST)
Advisory Committee
Multilingual Synchronization: LOD by NLP

- Imbalances between different sources are founded
  - Different degree of information

- DBpedia Infobox Example
  - Differences between different languages
    - En DBpedia: 43,641,010
    - Ko DBpedia: 646,187
Introduction: Goal

• Multilingual DBpedia Synchronization
  – Synchronization means **translation** between two languages + **information balancing** for their equivalent degree of information
Introduction: Goal

- Multilingual DBpedia Synchronization
  - Synchronization means **translation** between two languages + **information balancing** for their equivalent degree of information
Introduction

• Wikipedia
  – Supports over 270 languages
  – Allows cross-lingual navigation with *inter-language link*
  – Different quantity of data

• Goal
  – Synchronizing multilingual Wikipedia data to fill the gap between different languages & to acquire the integrated knowledge
Analysis of infoboxes in Wikipedia

• Analysis of infoboxes in Wikipedia
  – Infoboxes are manually created by authors
  – Many articles have no infoboxes and other articles contain infoboxes which are not complete

• Analysis of infoboxes in Multilingual Wikipedia
  – The inter-language linked articles do not use the same infobox template or contain different amount of information
  – Types of imbalances of infobox information btw different languages
    • Short Infobox
    • Distant Infobox
    • Missing Infobox
Type of infoboxes in Two Language Wikipedia

• **Short Infobox Type**
  – Use the same infobox template
  – but have a different amount of information

• **Distant Infobox Type**
  – Use different infobox templates
  – Due to
    • the different degrees of each Wikipedia communities’ activities
    • a matter from a different standpoint

• **Missing Infobox Type**
  – The infobox is discovered on only one side
Example of S-Infobox Type
Example of D-Infobox Type

Internet Explorer
From Wikipedia, the free encyclopedia

This article contains weasel words, vague phrasing that often accompanies biased or unverifiable information. Such statements should be clarified or removed. (February 2010)

Windows Internet Explorer (formerly Microsoft Internet Explorer; abbreviated to MSIE or, more commonly, IE), is a series of graphical web browsers developed by Microsoft and included as part of the Microsoft Windows line of operating systems starting in 1995. It has been the most widely used web browser since 1999, attaining a peak of about 95% usage share during 2002 and 2003 with IE5 and IE6.

That percentage share has since declined in the face of renewed competition from other web browsers – Mozilla Firefox most of all. Its usage share now sits at approximately 50% to 60% and is trending downward. Microsoft spent over $100 million a year on IE in the late 1990s, with over 1,000 people working on it by 1999. The latest release is Internet Explorer 8, which is available as a free update for Windows XP with Service Pack 2 or later, Windows Server 2003 with Service Pack 1 or later, Windows Vista, and Windows Server 2008, and is included with Windows 7 and Windows Server 2008 R2. Internet Explorer was originally going to be omitted from Windows 7 and Windows Server 2008 R2 in Europe, but Microsoft later reversed this decision.

Example Infobox for Windows Internet Explorer

<table>
<thead>
<tr>
<th>Infobox Windows component</th>
</tr>
</thead>
<tbody>
<tr>
<td>name = Windows Internet Explorer</td>
</tr>
<tr>
<td>logo = internet Explorer 7 Logo.png</td>
</tr>
<tr>
<td>screenshot = Internet Explorer 8.png</td>
</tr>
<tr>
<td>caption = Internet Explorer 8 in [[Windows Vista]]</td>
</tr>
<tr>
<td>included_with =</td>
</tr>
<tr>
<td>replaces =</td>
</tr>
<tr>
<td>replaced_by =</td>
</tr>
<tr>
<td>related_components = Versions: &lt;br /&gt;{Msieversions}</td>
</tr>
</tbody>
</table>

Key-Sun Choi, Linked Data for or by NLP
Example of M-Infobox Type

---

**Dennis Ritchie**

From Wikipedia, the free encyclopedia.

Dennis MacAlistair Ritchie (username: dmr, born September 9, 1941) is an American computer scientist notable for his influence on C and other programming languages, and on operating systems such as Multics and Unix. He received the Turing Award in 1983 and the National Medal of Technology in 1996. Ritchie was the head of Lucent Technologies System Software Research Department when he retired in 2007.

**Contents**

- Background

**Background**

Born in Bronxville, New York, Ritchie graduated from Harvard University with degrees in physics and applied mathematics. In 1967, he began working at the Bell Labs Computing Sciences Research Center.

**C and Unix**

Ritchie is best known as the creator of the C

---

**Dennis MacAlistair Ritchie**

**Dennis Ritchie (right) with Ken Thompson**

**Bom**

September 9, 1941 (age 68)

Bronxville, New York

**Fields**

Computer Science

**Institutions**

Lucent Technologies

Bell Labs

**Known for**

ALTRAN

B

BCPL

C

Multics

Unix

**Notable awards**

Turing Award

National Medal of Technology
Enriching Infobox using DBpedia Translation

• Basic processes
  – Extracting all tuples of infoboxes
  – Translating English tuples into Korean
  – Automatic complementing the Korean infoboxes

• Experimentation Result about DBpedia Translation
  – Source: English triples in DBpedia: 43M triples
  – Translated triples: 12M triples
    • Existing triples in Korean DBpedia: 485,000
  – Over 20 times triples are created by translation
Related Work

• DBpedia
  – A large, on-going, project which concentrates on the task of converting Wikipedia content into structured knowledge
  – The infobox extraction algorithm detect such template and recognizes their structure and save it in RDF triples

• OntoCloud
  – Automatic building the template ontology
    • Helpful resource to navigate template (infobox)
    • Suggesting the most relevant template with the article
Infobox Sync. Overview

• Infobox ≡ Fact sheet
  – Enriching Infobox from other language resources
    • Quantity Expansion
      – English 1,552,670 articles >>> Korean 43,444 articles
    • Quality Expansion
      – Based on different intentions, perspectives, etc.
  – Making further improvements
    • Information evolves over time
    • Adding temporal events (Fact Sheet Flow)
      – Extracting events from articles
Enriching Infobox from other language resources

• 3 types of Imbalances
  – Insufficient
  – Different
  – Missing


Extracting → Translating → Integrating
(Eliminating overlapping/ Resolving conflicts)
Ontology Mapping with WN

• 3 types of mapping approach
  – Based on Name similarity
    • mapping concepts which have same name or same meaning of definition
  – Based on Hierarchy similarity
    • Mapping concepts which be placed similar position of hierarchy
  – Based on Instance similarity
    • Mapping concepts which contain similar group of instance
Cultural Sync

• Hypothesis
  – X is a key fact in $L_1 \rightarrow X'$ should be a key fact in $L_2$
    • where $X'$ is a corresponding term to X in different language
      – Assumption
        » Inter-language links are accurate links to connect two pages about the same entity or concept in different languages

• Key facts come from the structured data such as:
  – Infobox
  – Category
  – Hyperlink text (than normal text)
Methodology of M-Sync

• Basic methodology
  – Infobox synchronization
    • Duplicate resolving & conflict resolving

• Issues
  – Multilingualism
  – Multiple viewpoints
  – How to evaluate
Automatic construction of initial DBpedia in other Languages
Automatic Construction of Korean initial DBpedia from English DBpedia

**Step 1:** Extract shared entities between English Wikipedia & Korean Wikipedia
* Shared entities = Interlingual connected entities in both languages

**Step 2:** Extract subset of English DBpedia with shared entities

**Step 3:** Convert \(<Title, Property, Value>\) format into \(<Title, Template, Property, Value>\)
* Example)
  - DBpedia triple:
  <Thomas Edison, wikiPageUsesTemplate, Template:Infobox_person>,
  <Thomas Edison, parents, Nancy Matthews Elliott>,
  <Thomas Edison, birthDate, 1847-02-11>
  - Converted quadruple:
  <Thomas Edison, Infobox_person, parents, Nancy Matthews Elliott>,
  <Thomas Edison, Infobox_person, birthDate, 1847-02-11>

**Step 4:** Extract \(<Template, Property>\) pairs and translate it into Korean using common dictionaries
* The words used in Template or Property are common words (easy to translate)

**Step 5:** Translate Value in English into Korean using WikiDictionary and Google translation API (except date, numeric value, …)
Modelling on influence links

• Example of links in multiple language Wikipedia
  • Different Wikipedia has different viewpoints and different concerns
  • Some links are newly added, some others are deleted by user in a temporal manner
Methodology of M-Sync

• Basic methodology
  – Infobox synchronization
    • Duplicate resolving & conflict resolving

• Issues
  – Multilingualism
  – Multiple viewpoints
  – How to evaluate
Summary: LOD by NLP

• Linked Open Data is based on the collectively generated content, e.g., Wikipedia.
• But there is information imbalance in multilingual articles
• NLP technology will assist this status to overcome the multilingual synchronization by
  – Template generation assist
  – Template maintenance of infobox
  – Linked Data Assistant for semantics of links
Collaboration in Asia

Organizational Efforts:
- Propose “Asian Forum on Multilingual (Semantic) Web”

Activities
- More Active Collaboration with EU and US projects [1]
- Training and Dissemination [2]
- “SIG-Standards” under AFNLP (Asia Federation of NLP) [3]
- EAFTERM (East-Asia Forum on Terminology)
  - Expanding to AsiaTerm
- Encoding scheme [4]

[1] EU-FP7 Iod2: Korean/Asian localization/Fusion to LOD (KAIST) http://lod2.eu
Is LOD for NLP?

Wikipedia as gazetteers for Named Entity Recognition
NER: Introduction

- Definition of Named Entity Recognition (NER):
  - To classify phrases that contain the names of persons, organizations and locations (Kim et al., 2003)
  - To classify every word in a document into some predefined categories and "none-of-the-above" (Zhou et al., 2002)

Example from [Ratinov et al., 2009]

SOCCER – [BLINKER][PER] BAN LIFTED. [LONDON][LOC] 1996-12-06 [Dutch][MISC] forward [Reggie Blinker][PER] had his indefinite suspension lifted by [FIFA][ORG] on Friday and was set to make his [Sheffield Wednesday][ORG] comeback against Liverpool on Saturday. [Blinker][PER] missed his club’s last two games after [FIFA][ORG] slapped a worldwide ban on him for appearing to sign contracts for both [Wednesday][ORG] and [Udinese][ORG] while he was playing for [Feyenoord][ORG].
Importance of “External Knowledge”

- Prior knowledge from outside resource is very helpful in determining the class of each recognized entity, as the following example shows:

**Example from Ratinov et al., 2009**

```
SOCER – [BLINKER]PER BAN LIFTED.
[LONDON]LOC 1996-12-06 [Dutch]MISC forward
[Reggie Blinker]PER had his indefinite suspension lifted by [FIFA]ORG on Friday and was set to make his [Sheffield Wednesday]ORG comeback against Liverpool on Saturday. [Blinker]PER missed his club’s last two games after [FIFA]ORG slapped a worldwide ban on him for appearing to sign contracts for both [Wednesday]ORG and [Udinese]ORG while he was playing for [Feyenoord]ORG.
```
Suggestion of gathering external knowledge:
From Wikipedia

- Wikipedia is:
  - Open, collaborative encyclopedia
  - Each entries of Wikipedia contains manually-tagged categories

- Can’t we use those categories to gather the external knowledge?
Collecting External Knowledge for NER from Wikipedia Category Structure

- Suggestion from [Ratinov et al., 2009]:
  1. Chooses some categories for each NE class
  2. Gather all the entities which have the predefined category tag
  3. Use the gathered entities as gazetteers

**World Bank**

From Wikipedia, the free encyclopedia

The **World Bank** is an international financial institution that provides loans to developing countries for capital programmes. The World Bank's official goal is the reduction of poverty. By law, all of its decisions must be guided by a commitment to promote foreign investment, international trade and facilitate capital investment.[3]

The World Bank differs from the World Bank Group, in that the World Bank comprises only two institutions: the International Bank for Reconstruction and Development (IBRD) and the International Development Association (IDA), whereas the latter incorporates these two in addition to three more: International Finance Corporation (IFC), Multilateral Investment Guarantee Agency (MIGA), and International Centre for Settlement of Investment Disputes (ICSID).


Gazetteer for “BANK” (ORG)

World Bank
Problems of “simply” collecting entities with category tags

- Many category tags simply represent “related” entity, not meaning `instanceOf` (subsumption) relation
  - Random test of small subset shows only 52.61% of category tags represent actual `instanceOf` relation (Choi et al., 2010)
  - In consequence, the resultant gazetteers contain many erroneous entities

Example of an entity which has category tag “Banks” but in fact is not a bank.

```
BPO security
From Wikipedia, the free encyclopedia

Information security has emerged as a significant concern for banks, mobile phone companies and other businesses that use call centers or business process outsourcing, or BPO. There have been instances of theft of personal data reported from call centers.

Britain's Financial Service Authority examined standards in India in April 2005 and the Banking Code Standards Board audited eight Indian call centres in 2006, handling more than a million calls per month from the UK. The BCSB report stated that "Customer data is subject to the same level of security as in the UK. High risk and more complex processes are subject to higher levels of scrutiny than similar activities onshore." [1]

India's NASSCOM has said that they take breach in security extremely seriously and will assist the police in their probe.

...```
Proposed Approach

- Rather than using the category tags directly, how about to….
  
  (1) Analyze the category tags
  
  (2) Gather “words” which represents what the entity is
  
  (3) Construct gazetteers using those “words”

So that,

(1) Remove erroneous data, and

(2) Produce gazetteers for any types of NE, even if the type is not one of those category tags.
### Proposed Approach

(1) Gathering upper category tag sets of each entity

<table>
<thead>
<tr>
<th>Entity</th>
<th>upper category tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France, Banks of France, Central banks, Economies by Country France,</td>
</tr>
<tr>
<td></td>
<td>Financial services companies of France, European Countries, Companies of France by</td>
</tr>
<tr>
<td></td>
<td>industry</td>
</tr>
</tbody>
</table>

(2) Analyzing upper category tags to get tokens which represents identity of each entity

<table>
<thead>
<tr>
<th></th>
<th>Identity-Representing tokens of “Banque de France”</th>
</tr>
</thead>
<tbody>
<tr>
<td>As Head</td>
<td>Bank, …</td>
</tr>
<tr>
<td>As Modifier</td>
<td>France, …</td>
</tr>
</tbody>
</table>

(3) Generating gazetteers based on the analysis result

Gazetteer for “BANK” (ORG)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>World Bank</td>
<td>Banque de France, …</td>
</tr>
<tr>
<td></td>
<td>…</td>
</tr>
</tbody>
</table>
“Identity-Representing” tokens of an entity

- What property should those tokens represent?
  - “essential property” of Guarino [2000]/ “intrinsic property” of Mizoguchi [2004]
  - Two concepts represent the same thing
A rigid property is a property that is essential to all its instances.

- Example:

  1. If \( x \) is an instance of \textit{PERSON}, it must be an instance of \textit{PERSON} in every possible world \( \rightarrow \) property \textit{PERSON} is rigid. (\(+R\))

  2. Non-rigid (\(-R\)) property is a property that is not essential to some of its instances.

     a. If \( x \) is an instance of \textit{STUDENT}, it could not be an instance of \textit{STUDENT} in some world \( \rightarrow \) property \textit{STUDENT} is not rigid (anti-rigid). (\(\sim R\))

     b. Semi-rigid (\(\neg R\)) is a property that is non-rigid but not anti-rigid (Ex: \textit{PERSON OR STUDENT})
Is it possible to recognize individual entities in the same world as the same (or different)? – If possible, the property has *identity condition (IC)*.

- Example:

1. Property *ENTITY*: does not have identity condition (-I)
2. Property *PERSON*: *SUPPLIES* identity condition (+O)
3. Property *STUDENT*: Carries identity condition (+I), but does not supply it (-O).
A property $\psi$ is externally dependent on property $\phi$ if, for all its instances $x$, necessarily some instance of $\psi$ must exist, which is not a part nor a constituent of $x$.

- Example:

  Property $PARENT$ is dependent on property $CHILD$
Ideal taxonomy structure proposed by Guarino

Backbone Taxonomy

Categories (-O-I+R+D)

Attributions (-O-I~R-D) (-O-I→R)

Top Types

Types & Quasi-Types (+O+I+R) (-O+I+R)

Mixins (-O+I→R)

Formal Roles (-O-I~R+D)

Material Roles (-O+I~R+D)

Phased Sortals (-O+I~R-D)
Mizoguchi’s definition of “instanceOf”

- **Definition of “class”:**
  
  A thing which is a conceptualization of a set \( X \) can be a class, iff each element \( x \) of \( X \) belongs to the class \( X \) iff the “**intrinsic property**” of \( x \) satisfies the intensional condition of \( X \)

- **Definition of “**\( x \) instanceOf \( X \)”**: 
  
  \(<x \text{ instanceOf } X>\) holds if and only if the intrinsic property of \( x \) satisfies the intensional condition of \( X \), where \( x \) is an entity and \( X \) is a class.

- **“Intrinsic property”** here is the same concept as **essential property** of Guarino’s paper (+R).
Conclusion: Targeted Property

- We will try to make the extracted tokens to represent “essential property/intrinsic property” of an entity:
  - Following the definition of Guarino’s backbone taxonomy, and
  - The definition of Mizoguchi’s `instanceOf`-relation
① Each category tag expresses the property of category itself
- Since Wikipedia is a “collaborative” encyclopedia

② Assumption of Category Construction
- Classical view of Category: Discrete entities characterized by a set of properties shared by their members
- Intrinsic property: Which is essential to a thing.
- Thus, intrinsic property will be used much more frequently than other properties when defining categories
Proposed Approach: Intrinsic Property-based Approach

What is Intrinsic Property?

✓ Intrinsic property of a thing is, a property which is essential to the thing, and it loses its identity when the property changes. [Guarino et al, 2000]

➢ Ex: The intrinsic properties of ‘Pioneer 11’ are:
  - Spacecraft, Jupiter, Saturn, …

How could we get them for a category node?

(1) Categories are: discrete entities characterized by a set of properties shared by their members.
(2) Since intrinsic properties represent identity of a thing, intrinsic properties will be shared by the members of categories frequently.
Assumptions for property extraction (2/2)

Assumption of Co-occurrence

- If a token $A$ co-occurs with other token $B$ which represents identity of entity $C$, then there is a high probability that token $A$ also represents identity of entity $C$.

- Co-occur here means: Co-occurrence of two tokens, one as head of a category tag, and the other as modifier of that category tag.

---

**Diagram:**

- **Central Banks**
  - Category Tag
  - Represents identity
  - Co-occurs

- **Banque de France**
  - Category Tag
  - Represents identity
  - entity

- **Central Banks**
  - Category Tag
  - Represents identity

- **Banque de France**
  - Category Tag
  - Represents identity
  - entity
Construction of Modifier Graph

- Based on the assumption, “Modifier graph” for each entity is created.

- Modifier graph is a directed graph, defined as
  - A vertex: Represents a token from upper category tags
  - An edge: one vertex modifies other side vertex in one of the category tags
  - Weight of an edge: Number of emergence of two vertices as modifier–head word relation inside the upper category tags

- Goal of Graph Analysis:
  - Finding high-scored vertex, which is considered to represent the identity of an entity with high reliability
Gathering Upper category tags

- To collect enough information, enough number of upper category tags should be collected.
- 25+ upper category tags of an entity are gathered to construct a modifier graph for the entity.

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France, Banks of France, Central banks,</td>
</tr>
<tr>
<td></td>
<td>Economies by Country, France, Financial services,</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
<tr>
<td></td>
<td>European countries, Companies of France by industry</td>
</tr>
</tbody>
</table>

Diagram:
- Target entity: Banque de France
- Connected to: Economies by Country, France, Financial services companies of France, Central banks, Companies of France by industry, European countries.
## Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France</td>
</tr>
<tr>
<td></td>
<td>Banks of France</td>
</tr>
<tr>
<td></td>
<td>Central banks</td>
</tr>
<tr>
<td></td>
<td>Economies by Country</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Financial services</td>
</tr>
<tr>
<td></td>
<td>companies of France</td>
</tr>
<tr>
<td></td>
<td>European countries</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
</tbody>
</table>
## Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td><em>Economy of France</em></td>
</tr>
<tr>
<td></td>
<td>Banks of France</td>
</tr>
<tr>
<td></td>
<td>Central banks</td>
</tr>
<tr>
<td></td>
<td>Economies by Country</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Financial services companies of France</td>
</tr>
<tr>
<td></td>
<td>European countries</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
</tbody>
</table>

**Economy**
### Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France</td>
</tr>
<tr>
<td></td>
<td>Banks of France</td>
</tr>
<tr>
<td></td>
<td>Central banks</td>
</tr>
<tr>
<td></td>
<td>Economies by Country</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Financial services</td>
</tr>
<tr>
<td></td>
<td>companies of France</td>
</tr>
<tr>
<td></td>
<td>European countries</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
</tbody>
</table>
## Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France</td>
</tr>
<tr>
<td></td>
<td>Banks of France</td>
</tr>
<tr>
<td></td>
<td>Central banks</td>
</tr>
<tr>
<td></td>
<td>Economies by Country</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Financial services companies of France</td>
</tr>
<tr>
<td></td>
<td>European countries</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
</tbody>
</table>
Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France</td>
</tr>
<tr>
<td></td>
<td>Banks of France</td>
</tr>
<tr>
<td></td>
<td>Central banks</td>
</tr>
<tr>
<td></td>
<td>Economies by Country</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Financial services companies of France</td>
</tr>
<tr>
<td></td>
<td>European countries</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
</tbody>
</table>

Key: Sun Choi, Linked Data for or by NLP
Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France</td>
</tr>
<tr>
<td></td>
<td>Banks of France</td>
</tr>
<tr>
<td></td>
<td>Central banks</td>
</tr>
<tr>
<td></td>
<td>Economies by Country</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Financial services companies of France</td>
</tr>
<tr>
<td></td>
<td>European countries</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
</tbody>
</table>

Economy  ➔  France

Key-Sun Choi, Linked Data for or by NLP
Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France</td>
</tr>
<tr>
<td></td>
<td>Banks of France</td>
</tr>
<tr>
<td></td>
<td>Central banks</td>
</tr>
<tr>
<td></td>
<td>Economies by Country</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Financial services companies of France</td>
</tr>
<tr>
<td></td>
<td>European countries</td>
</tr>
<tr>
<td></td>
<td>Companies of France by industry</td>
</tr>
</tbody>
</table>

Key:
- Economy
- France
- Bank
### Modifier Graph for ‘Banque de France’

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France, Banks of France, Central banks,</td>
</tr>
<tr>
<td></td>
<td>Economies by Country, France, Financial services</td>
</tr>
<tr>
<td></td>
<td>companies of France, European countries, Companies of</td>
</tr>
<tr>
<td></td>
<td>France by industry</td>
</tr>
</tbody>
</table>
## Modifier Graph for ‘Banque de France’

### Graph Diagram

- **Economy** → **France**
  - 1
- **France** → **Bank**
  - 1

### Table

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Upper Category Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque de France</td>
<td>Economy of France, Banks of France,</td>
</tr>
<tr>
<td></td>
<td>Central banks, Economies by Country,</td>
</tr>
<tr>
<td></td>
<td>France, Financial services companies of</td>
</tr>
<tr>
<td></td>
<td>France, European countries, Companies</td>
</tr>
<tr>
<td></td>
<td>of France by industry</td>
</tr>
</tbody>
</table>
Modifier Graph for ‘Banque de France’

### Target Entity
- Banque de France

### Upper Category Tags
- Economy of France
- Banks of France
- Central banks
- Economies by Country
- France
- Financial services companies of France
- European countries
- Companies of France by industry

---

![Diagram](image-url)

### Diagram Details
- **Economy**: 1
- **Country**: 1
- **European**: 1
- **France**: 1
- **Bank**: 1
- **Company**: 1
- **Financial**: 2
- **Service**: 1
- **Industry**: 1
Actual Modifier Graph for ‘Banque de France’
Goal:

(1) To find out tokens which represent identity of the entity by itself along
   - Ex. “Bank” for entity “Banque de France”

(2) To find out tokens which shows identity of the entity, but cannot be used by itself along
   - Ex. “Central” for entity “Banque de France”
Algorithm for graph analysis: HITS

- HITS link analysis algorithm is a well-known page ranking algorithm
- “Authority” score measures importance of page contents, and “Hub” score measures importance of links from the page
- Authority Update Rule:
  \[ \text{Authority}(V_i) = \sum_{V_j \in \text{In}(V_i)} e_{ji} \cdot \text{Hub}(V_j) \]
- Hub Update Rule:
  \[ \text{Hub}(V_i) = \sum_{V_j \in \text{Out}(V_i)} e_{ij} \cdot \text{Authority}(V_j) \]
- **Algorithm**

1. Initialize hub/authority score of each node to 1
2. Update authority score
3. Update hub score
4. Normalize scores so that the squared sum of every authority score and the squared sum of every hub score to be 1.
5. Iterate from step 2 as necessary.
Analysis Result for modifier graph of: “Banque de France”

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Banque de France</th>
</tr>
</thead>
</table>
| **Tokens representing identity of the entity, and can stand alone** | Bank (0.8232)  
Economy (0.3427)  
Government (0.3203)  
Politics (0.1780)  
Banking (0.1725) |
| **Tokens representing identity of the entity, but cannot stand alone** | France (0.6411)  
Central (0.6335)  
Country (0.3848)  
Europe (0.0656)  
European (0.0554) |

• Full analysis result is available at:  
http://semanticweb.kaist.ac.kr/home/index.php/IntrinsicProperty
Generating Gazetteers with keywords

Steps:

1. Get “keywords” for the targeted gazetteers
2. Calculate relevance score of each entity to the keywords, using their analyzed results (tokens representing the entity’s identity, with their relevance score)
3. Get entities whose relevance score to the keyword is over a predefined threshold $\mu$. 

Key-Sun Choi, Linked Data for or by NLP
Example of making Gazetteers for “ORG”

Keyword for ORG: \textbf{Banks}, \mu = 0.1

Calculation of relevance score of entity “Banque de France” to keyword “Banks”: squared sum of relevance of each identity-representing tokens which is contained by the keyword

<table>
<thead>
<tr>
<th>Target Entity</th>
<th>Banque de France</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens representing identity of</td>
<td>Bank (0.8232)</td>
</tr>
<tr>
<td>the entity, and can stand alone</td>
<td>Economy (0.3427)</td>
</tr>
<tr>
<td></td>
<td>Government (0.3203)</td>
</tr>
<tr>
<td></td>
<td>Politics (0.1780)</td>
</tr>
<tr>
<td></td>
<td>Banking (0.1725)</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Tokens representing identity of</td>
<td>France (0.6411)</td>
</tr>
<tr>
<td>the entity, but cannot stand</td>
<td>Central (0.6335)</td>
</tr>
<tr>
<td>alone</td>
<td>Country (0.3848)</td>
</tr>
<tr>
<td></td>
<td>Europe (0.0656)</td>
</tr>
<tr>
<td></td>
<td>European (0.0554)</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Relevance (Banque de France, Bank) = (0.8232)^2
= 0.6777 > \mu = 0.1
∴ Banque de France is included in the gazetteer for ORG
To figure out the effect of our new gazetteers, LBJNER, which is one of the state-of-the-art open NER tagger is used:

- It uses gazetteers from Wikipedia, which is simple collection of entities with some predefined category tags
- We collected our gazetteers with the exact same category tags as our keywords
- LBJNER: (Ratinov et al., 2009)
To figure out the effect of our new gazetteers, LBJNER, which is one of the state-of-the-art open NER tagger is used:

- We experimented using LBJ’s baseline system (LBJ-BASELINE).
- Test is carried out on CoNLL03 test data.
- Phrase-level precision/recall/F-measure is measured.
To gather more information, redirect information is gathered to more populate the gazetteer.

Gazetteer for “BANK” (ORG)

World Bank
Banque de France
Banque de france
...

Banque de france
From Wikipedia, the free encyclopedia
Redirect page

Banque de France
Comparison of the systems: Effect of the new Gazetteers

- $\mu = 0.02$

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBJ-BASELINE</td>
<td>84.51</td>
<td>82.81</td>
<td>83.65</td>
</tr>
<tr>
<td>LBJ-BASELINE + LBJ-WikiGazetteers</td>
<td>87.86</td>
<td>86.58</td>
<td>87.21</td>
</tr>
<tr>
<td>LBJ-BASELINE + Our Gazetteers</td>
<td>88.05</td>
<td>86.63</td>
<td>87.34</td>
</tr>
</tbody>
</table>

- A little increase in F1-score.
- Although many more experiments are required with different parameters, new method shows some possibility to improve the performance.
- Since LBJ-WikiGazetteers are already highly tuned for NER of CoNLL dataset, we need some different test data and different domain, to fully recognize the full power of the proposed method
Summary: LOD for NLP

- For Category tags in collectively generated content, e.g., Wikipedia,
  - They have no taxonomy.
- Intrinsic property of each entry of Wikipedia can be extracted from its category tags.
- The resulting focused tokens in modifier graph will be used for their NER gazetteer.
- All of extracted tokens will contribute to a taxonomy – that will supplement the complementary parts of already-made word nets.
Concluded Questions

- Does Wikipedia and LOD replace/complement WordNet?
  - By converting the Collectively Generated Categories to enrich the Synsets of WordNet?
  - IS-A Construction:

- What is the target of Sense Evaluation?
  - Template finding in Infobox
  - Selection task in Wikipedia Disambiguation page

- Approximation from Intrinsic Properties to Frequency-based algorithm

- Question Answering?
  - “Who are researchers working on NLP and Logic topics in East Asian countries?”
  - ...

- NLP classic vs. NLP with Linked Data?
CoreNet
Korean–Chinese–Japanese Wordnet with shared concept system
Contents

- Overview
- Structure
- Conceptual system
- Korean Wordnet
- Chinese Wordnet
- Process and Consideration
- Conclusion
Overview

- CoreNet
  - Concept based Multilingual Wordnet
    - Korean-Japanese-Chinese are aligned.
  - Accompanied with syntactic structure for verbs and adjectives

- Purpose
  - Resolving semantic ambiguity in the language processing

- Method
  - By the semi-automatic method utilizing a monolingual MRD and an existing wordnet
Process for CoreNet

1. Large scale corpus
2. Dictionary “Urimal”
3. Manual Selection of Korean general vocabulary
4. First mapping of word sense to corresponding concept
5. Second mapping (Selection of candidates among results of first mapping)
6. Usages extracted from large-scale corpus
7. Specialist’s manual selection
8. Web based environment Specialist’s selection
9. Noun Wordnet
10. Verb and Adjective Wordnet
11. Korean Wordnet
CoreNet through Lexical map integrated browsing tool

Lexical Map sorted by ENTRY

Lexical Map by Concept system

Monolingual Paper dictionary

Semantic category, or concept
Three main components

- Shared concept system
- Korean Wordnet
- Chinese Wordnet
Shared concept system

- Concept system
  - Originated from NTT lexical hierarchy

친구/지인 | 友・なじみ | 朋友/熟人: 124
  - 친구 | 友人 | 朋友: 125
  - 지인 | 知人 | 熟人: 126
  - 연인 | 恋人 | 情人: 127
  - 남자 연인 | 恋人（男） | 情人(男子): 128
  - 여자 연인 | 恋人（女） | 情人(女子): 129
Korean Wordnet

- Concept system for Korean
  - 2794 concept nodes

- Korean word senses assigned to the corresponding concept nodes:
  - 23,823 nouns (56,523 senses)
  - 1,757 verbs (4,717 senses)
  - 804 adjectives (1,392 senses)

- Syntactic structure of Korean verbs and adjectives
- List of arguments for each verb and adjective sense
- Japanese translations
<table>
<thead>
<tr>
<th>Pre-subdivision</th>
<th>Post-subdivision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Place</strong></td>
<td><strong>Arrive/ Intransitive verb /0/1, Arrive/ Intransitive verb /1/2, Tread/ Transitive verb /0/3, Arrive/ Intransitive verb /1/1, Approach/ Intransitive verb /0/2</strong></td>
</tr>
<tr>
<td><strong>Situation</strong></td>
<td>***1Come/ Intransitive verb /0/3, Approach/ Intransitive verb /1/0, *1Meet/ Intransitive verb /0/1, *1 Meet/ Transitive verb /1/1, Approach / Intransitive verb /0/0, *1Encounter/ Intransitive verb /0/4, *1Meet/ Transitive verb /0/2, *1Greet/ Transitive verb /1/1, *1 Receive / Transitive verb /1/2, *1 Be/ Intransitive verb /0/0, *1Hit/ Transitive verb /0/7, <em>1Meet/ Transitive verb /0/4, <em>1 Face/ Transitive verb /1-1/0</em></em></td>
</tr>
<tr>
<td><strong>Extent</strong></td>
<td>***1Arrive/ Intransitive verb /1/3, Come/ Intransitive verb /2/4, Climb/ Intransitive verb /0/2, *1Arrive/ Intransitive verb /1/4, <em>1Grow/ Intransitive verb /2-1/0, <em>1Eat/ Transitive verb /0/7,</em></em></td>
</tr>
<tr>
<td><strong>Position</strong></td>
<td>**<em>1Stand/ Intransitive verb /1/4 <em>1Sit/ Intransitive verb /0/4</em></em></td>
</tr>
</tbody>
</table>

| **Separation**  | **Divide/ Transitive verb /2/2, Divide/ Intransitive verb /1/2, Divide/ Intransitive verb /1/0, Divide/ Transitive verb /0/1, Divide / Intransitive verb /0/0, Open/ Transitive verb /2/1, Open/ Intransitive verb /0/1, Open/ Intransitive verb /0/2, Push/ Transitive verb /0/1, Tear/ Transitive verb /5/1, Rip/ Transitive verb /5/3, Split/ Transitive verb /0/0, Split/ Intransitive verb /0/0, Tear/ Transitive verb/0/1, Split/ Transitive verb /0/2, Split/ Intransitive verb /0/0, Burn/ Transitive verb /4/1, Burn/ Transitive verb /4/1, Filter/ Transitive verb /1/0, Unload/ Transitive verb /1-1/6, Remove/ Transitive verb /3/2, Deduct/ Transitive verb /2/1, Sweep / Intransitive verb /2/1, Drop/ Transitive verb /0/2, Fall/ Intransitive verb /0/2, Separate/ Transitive verb /0/1, Remove/ Transitive verb /0/9, Leave/ Transitive verb /7/1, Open/ Intransitive verb /1/1, Remove/ Transitive verb /0/1, Leave/ Transitive verb /0/0, Remove/ Transitive verb /0/3, Leave/ Transitive verb /1-1/4, Undo/ Transitive verb /0/2** |
| **Having it**   | **Filter/ Transitive verb /1/0, Unload/ Transitive verb /1-1/6, Remove/ Transitive verb /3/2, Deduct / Transitive verb /2/1, Sweep / Intransitive verb /2/1, Drop/ Transitive verb /0/2, Fall/ Intransitive verb /0/2, Separate / Transitive verb /0/1, Remove / Transitive verb /0/9, Leave/ Transitive verb /7/1, Open / Intransitive verb /1/1, Remove / Transitive verb /0/1, Leave / Transitive verb /0/0, Leave / Transitive verb /0/3** |
| **Having it**   | **Separate / Transitive verb /0/6 Remove / Transitive verb /0/7 Leave/ Transitive verb /2/0 Leave /Transitive verb/1/0 Split/ Transitive verb /3/2** |
| **Others**      | **Leave/ Transitive verb /1-1/4 Undo/ Transitive verb /0/2** |
Korean-Japanese Verb Translation - corpus and concept-based

가다 0 [VI]

(1) 2174#가기|行き|出行
   N1-subj N2-dative
   N1[5#人間] N2[5#人間]-行く
   N1-subj N2-locative
   N1[5#人間] N2映画管[409#劇場]-行く

(3) 1550#전달|伝達|传达
   N1 기별 [1544#통신/소식/etc|通信・音信等|通信/消息]
   N2 집 [447#거주시설|居住施設|居住设施]-傳わる
**Web based tool for constructing caseframe and argument list**

![Web based tool screenshot](image_url)

<table>
<thead>
<tr>
<th>Web based tool for constructing caseframe and argument list</th>
<th>'말다'의 둥배 (현재: 77 / 전체: 94)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11: 몫맞이 알고 속맞이 돋 쌓는 것은 모두가 과정이 갈수록 이해할 수 있다.</td>
<td>미전</td>
</tr>
</tbody>
</table>

| 달다/형용사/6/1 | 줄이나 사람의 말과 같이. |
| 달디/형용사/6/2 | 맵네 비싼 말기에 많이 좋다. |
| 달다/형용사/6/3 | 마음에 담았다고 가깝다. |
| 달다/형용사/6/4 | 안타깝거나 소마조마하여 마음이 몫시 조금해져다 |

<table>
<thead>
<tr>
<th>의미/말다/형용사/</th>
<th>정의문</th>
</tr>
</thead>
<tbody>
<tr>
<td>입력하세요</td>
<td></td>
</tr>
</tbody>
</table>

| '말다/형용사/6/1'의 문형 목록 | |
|-----------------------------| |
| '말다/형용사/6/1' | N1 | N2 | N3 | N4 | 달다 |

참고사항 | 입력 |

<table>
<thead>
<tr>
<th>'N101/말다/형용사/6/1'의 적절부</th>
<th>비고</th>
</tr>
</thead>
<tbody>
<tr>
<td>과일/명사/2/0</td>
<td>염음</td>
</tr>
<tr>
<td>맡/명사/1/1</td>
<td>염음</td>
</tr>
<tr>
<td>별/명사/2/0</td>
<td>염음</td>
</tr>
<tr>
<td>별매/명사/1/1</td>
<td>염음</td>
</tr>
<tr>
<td>묤질/명사/0/2</td>
<td>염음</td>
</tr>
<tr>
<td>초유행/명사/0/0</td>
<td>염음</td>
</tr>
<tr>
<td>포도/명사/2/0</td>
<td>염음</td>
</tr>
<tr>
<td>흠치/명사/0/0</td>
<td>염음</td>
</tr>
</tbody>
</table>

참고사항 | 입력
# CoreNet (Korean Nouns)

The CoreNet (CoreNet) project is a comprehensive annotator for Korean nouns. It provides a list of key concepts and their relationships, which are crucial for understanding the context and meaning of Korean nouns. The project is particularly useful for researchers and students working on Natural Language Processing (NLP) tasks in Korean.

The CoreNet database includes annotations for various types of nouns, such as nouns, adjectives, and verbs, which are essential for developing effective NLP systems. The annotations are based on a combination of manual analysis and machine learning algorithms, ensuring accuracy and reliability.

The project also includes tools for analyzing and visualizing the data, making it easier to explore and understand the relationships between different nouns. These tools can be used to identify patterns and trends in the data, as well as to develop new NLP models and applications.

Overall, the CoreNet project is an important resource for anyone working with Korean language data, offering a wealth of information and tools for exploring the intricacies of the Korean language.
CoreNet (Korean Verbs)
CoreNet (Korean Adjectives)
Chinese Wordnet

- Concept system in Chinese
  - 2710 concept nodes

- Chinese word senses assigned to the corresponding concept nodes
  - 34,041 nouns (20647-28932 senses)
    - from “Peking University Chinese Grammar Information Dictionary (现代汉语语法信息词典)”.
  - 288 verbs (765 senses)
    - from “新现代汉语八百800词 (吕叔湘, 2002)”
  - 80 adjectives (119 senses)

- Syntactic structure of Chinese verbs and adjectives
- List of arguments for each verb and adjective sense
- Korean translations
Chinese-Korean Concept-based Noun

- **技艺**
  1) [jìyì] (N) 기교 [2506 기예|技芸|技艺<才能>]
  2) [jìyì] (N) 기술 [2506 기예|技芸|技艺<才能>]
  3) [jìyì] (N) 기예 [1670 재주<오락>|芸|技艺<娱乐>]
  4) [jìyì] (N) 솜씨 [2502 능력|能力|能力] [2506 기예|技芸|技艺<才能>]

- **计策**
  1) [jìcè] (N) 계략 [1036 안|案|方案<逻辑/意味等>] [1461 책략|策謀|策略]
  2) [jìcè] (N) 계책 [1036] [1461]
  3) [jìcè] (N) 솔책 [1035 방법|方法] [1461 책략|策謀|策略]

- **计程车** [jìchéngché] (N) 택시 [988 탈것(본체(땅))|乗り物(本体(移動(陸圏)))|运送工具(本体(移动(陆路)))]

- **计程仪** [jìchéngyì] (N) 측정기 [962 기계|機械|机械]

- **计划经济** [jìhuàjīngjì] (N) 계획경제 [1168 경제제도|制度(経済)|制度(经济)]
Chair [897]

- 安乐椅(안락의자) 按摩椅((의자) 宝座(옥좌) 长凳(벤치) 长椅(벤치) 方凳((의자) 扶手椅(휠체어) 交椅(의자) 靠背(등받이) 靠背椅(의자) 靠椅(의자) 罗圈椅((의자) 马扎((의자) 圈椅((의자) 软椅((의자) 沙发(소파) 沙发椅(소파) 睡椅((소파) 太师椅((의자) 躺椅(안락의자) 藤椅(등의자) 条凳(벤치) 土墩((의자) 椅背(등받이) 椅子(의자) 转椅(안락의자 회전의자)
Chinese–Korean Verb Translation

摘

1. [gǎo] [VT] 마련하다 (prepare, reserve)
   [N1]: 他(23,48) [V]: 搞 [N3]: 票(932)|电视机(970)

2. [gǎo] [VT] 만들다 (make)
   [N1]: 他(23,48) [V]: 搞 [Aux]: 了 [N3]: 方案(1036,1436,1460,1113)

3. [gǎo] [VT] 하다 (do)
   [N1]: 他(23,48) [V]: 搞 [N3]: 设计 | 施工 | 生产 | 工作

4. [gǎo] [VT] 경영하다 (manage)
   [N1]: 他(23,48) [V]: 搞 [N3]: 工厂(439)

5. [gǎo] [VT] 맡다 (take in charge of)
   [N1]: 他(23,48) [V]: 搞 [N3]: 总务(326)

6. [gǎo] [VT] 찾다 (search)
   [N1]: 他(23,48) [V]: 搞 [N3]: 对象(74)

7. [gǎo] [VT] 맺다 (form, contract)
   [N1]: 他们(25,2606) [V]: 搞 [N3]: 关系(1684,2444)

8. [gǎo] [VT] 구하다 (buy, look for)
   [N1]: 他(23,48) [V]: 搞 [N3]: 材料(769)

9. [gǎo] [VT] 쓰다 (use, adopt)
   [N1]: 他(23,48) [V]: 搞 [N3]: 花样(1461,1035,1009)
## CoreNet (Chinese)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>地区</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>人地系</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>地理</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>地球</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>大地</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>国际</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>区域</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Example

查找这段文字，“地”，查到的个数是 7 个。

- [467]地理
- [466]人地系
- [465]国际
- [464]行政区划
- [463]国际
- [462]区域
- [461]地理

- [460]人地系
- [459]区域

- [458]地理
Difficulties

- Chinese word - Korean compound word or syntagma
  - 学棍 xúegùn - 학교 깡패 (hooligan in the school)
  - 杏干 xìnggàn - 말린 살구 (dried apricot?)

- No corresponding Korean translation
  - 手感 (shǒugǎn) - 손으로 만져서 얻은 감각
  - 省优 (shěngyou) - 성에서 정한 우수한 것
  - 作陪 (zuówéi) - 함께 다니면서 안내도 하고 여러 가지로 일을 하는 것
Principles

- CoreNet has been constructed by the following principles
  - Word sense mapping to concept
  - Corpus based
  - Multi-lingualism
  - Mono-concept system for multi-languages
Sense Mapping to Concept

- Every sense of words in the dictionary is mapped to corresponding concepts.
  - School
    - Place/Organization/Building

- Each predicate has its own argument structure. Go “gada” has argument structures with concepts and Japanese translation as follows:
  - Going([human, mammal, vehicle]=subj) iku 行く
  - Learning ([human]=subj, [teacher]=dat), iku 行く
  - Delivery ([information]=subj, [human]=dat), tutawaru 伝わる
  - Progress ([time]=subj), sugiru
  - Continuation ([relation]=subj, [year]=obj), tuduku
  - Enthusiasm ([gaze]=subj, [girl]=dat), iku 行く
  - Sweep ([emotion]=subj), kieru
Corpus based Usage

- The set of vocabulary and their senses are extracted from KAIST corpus.
  - Going ([horse/mammal, bus/vehicle]=subj)
  - *Horse* and *bus* are extracted terms from corpus and *mammal* and *vehicle* are concept names mapped from words *horse* and *bus*.
  - This fact causes the more specified sense categorization than those of dictionaries.
Multi-lingualism

- All of concepts are aligned among three languages: Japanese, Korean, and Chinese.
- All of words (nouns and predicates) of three languages are categorized into one common concept hierarchy.
- Verbs of three languages are also linked each other based on senses and concepts.
  - qù (去)
  - going - gada (가다)
  - delivery - boneda (보내다)
  - exclusion - eobsaeda (없안다)
Considerations

- **Multiple inheritance**
  - One sense of a word is assigned to the multiple numbers of concepts.
    - School
    - location, organization, educational facility

- **Verbal noun**
  - A word in verb is assigned to concepts after it is transferred to its noun form.
    - *write* – writing
    - be wise – wisdom

- **Concept splitting**
  - Whenever we found the inconsistency among nodes of concepts, a node may be added.
    - body – three subconcepts (arm, leg, head)
    - But a word back?
Conclusion

- What difficulties must be overcome?
- Difference should be kept.
  - Difference
    - of mentality
    - of culture
    - of lexical system
CoreNet + WordNet Mapping

Semantic Web Research Center
@KAIST
Introduction of CoreNet

• CoreNet (KAIST)
  – Concept network of word senses for the Korean, Chinese, and Japanese languages
  – Its hierarchy originated from NTT Goi-Taikei concept hierarchy
  • For Korean:
    – 31,384 general words
    – total of 62,632 senses of them are linked to one or more concepts in CoreNet concept hierarchy
      » 2,937 high-level concepts mainly taxonomically organized into 12 depths
CoreNet + WordNet

• Goal
  – Linking semantic categories of CoreNet to WordNet
  – KorLex-based indirect-linking
    • Since CoreNet concepts are described in non-English languages, linking CoreNet to WordNet requires a language translation process
    • To alleviate such difficulties in Korean-to-English (K-E) translation and to increase recall of WordNet equivalents, we pursue to indirectly associate CoreNet with WordNet through
KorLex (Korean WordNet)

• KorLex is WordNet-referenced Korean WordNet
  – It has been developed since 2004
  – It contains about 130,000 synsets and 150,000 word senses for nouns, verbs, adjectives, adverbs, and classifiers
  – It is constructed by using the synset IDs of WordNet 2.0

• Although KorLex does not cover all WordNet synsets or vice versa, overlapped synsets between KorLex and WordNet contribute not only to reduce K-E translation problems but also to recall most synonymous K-E translations
Linked Relation types in WordNet

• 7 relation types are used as linking between a CoreNet concept and a WordNet synset
  – synonymy, hypernymy, hyponymy, troponymy, proper inclusion, presupposition, cause
  – For instance, when a CoreNet concept corresponds synonymously to WordNet synset
    {department_store, emporium}, its mapping is represented as follows
      • {department_store, emporium\textsuperscript{noun:03061806}\textsuperscript{^synonym\textsuperscript{3}}}
NLP-base Smart Social Calendar

- From E-mail, SNS and Web, to extract meeting and appointment information between a Mobile User and Other Related Users; and to notify them smartly.

Smart Email-Calendar Framework

Calendar Template

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Time</td>
<td>미팅의 시작시간(완료)</td>
</tr>
<tr>
<td>End Time</td>
<td>미팅의 종료시간(완료)</td>
</tr>
<tr>
<td>Loc</td>
<td>미팅의 장소(완료)</td>
</tr>
<tr>
<td>Participant</td>
<td>미팅에 참석하는 참여자</td>
</tr>
<tr>
<td>Preparation</td>
<td>미팅에 필요한 준비물</td>
</tr>
<tr>
<td>Host</td>
<td>미팅의 주최자</td>
</tr>
<tr>
<td>Title</td>
<td>미팅의 제목</td>
</tr>
<tr>
<td>Subject</td>
<td>미팅의 주제</td>
</tr>
<tr>
<td>Presenter</td>
<td>미팅의 발표자</td>
</tr>
<tr>
<td>Vehicle</td>
<td>미팅 장소까지 이용할 이동수단</td>
</tr>
<tr>
<td>Nearby Info</td>
<td>미팅 목적지 주변에 있는 이정표 정보</td>
</tr>
<tr>
<td>Target Audience</td>
<td>어떤 사람을 대상으로 하는 미팅인가</td>
</tr>
<tr>
<td>Contact</td>
<td>Email, 휴대폰번호 등 연락처 정보</td>
</tr>
</tbody>
</table>
Open Knowledge Convergence

Human Interface

Open Knowledge Service Framework

Data Catalog

Multilingual Link

Ontology

Data Storage

Trustfulness

Creativeness

Version History

Tagging

Location

Analytics

Web

Authoring (Data Add)

Retrieval

Preview

Visualization

Social Network

API

Key-Sun Choi, Linked Data for or by NLP
International Asian Summer School on Linked Data (IASLOD 2012), 13-17/Aug/2012

http://semanticweb.kaist.ac.kr/2012lodsummer/

IASLOD 2012 offers a broad and intensive series of lectures at different levels on Linked Data, ranging from Natural Language Processing to Web Engineering. The Linked Data methodology is a lightweight approach to facilitate the transition from the current web of unstructured documents to a web of structured data. Due to the large availability of Linked Data tools, industrial applications, and knowledge bases on one side, and the growth of the R&D community on the other side; the Linked Data paradigm has become a crucial building block of the Web architecture.
Key-Sun Choi
kschoi@kaist.ac.kr
A Creative Web with Natural Language

- Prof. Jae-Sung Lee (morph)
- Dr. Park, Jung-Yeol (parser)
- Dr. Yoon, Yong-Un (db)
- Dr. Martin Rezk (onto)
- Prof. Yi, Mun Y. (lod2.eu)
- Prof. Koh, In-Young (lod)
- Prof. Tony Veale (UCD)

http://semanticweb.kaist.ac.kr