Morpheme-based Annotation Tool for Korean Text

Andrew Cattle, Munhyong Kim, Hyopil Shin
Computational Linguistics Lab
Seoul National University
Table of Contents

• Introduction to Korean Sentiment Corpus
• Overview of Korean Language and Writing
• Overview of Existing Annotation Tools
• Design and Implementation
• Future Directions
Korean Sentiment Corpus

- Annotates subjectivity expressions in Korean text
- Morpheme-based annotations
  - Due to Korean’s agglutinative nature
- Consists of 332 newspaper articles taken from the fine-grained morphologically analyzed Sejong Corpus
  - Covers life, society, culture, and politics
- Uses Korean Subjectivity Markup Language (KSML) (Shin et al, 2012a, 2012b)
KSML Overview

- Three types of tag: SEED, SUBJECTIVE, OBJECTIVE
- SEED tags represent subjective statements smaller than a sentence. Contain:
  - Nested-source (the source of the subjectivity)
  - Type (how subjectivity is expressed)
  - Subjectivity-type (specific type of subjectivity)
  - Target (object the subjectivity is directed towards)
- SUBJECTIVE and OBJECTIVE tags denote when an entire sentence is subjective/objective (respectively)
나는 철수를 좋아한다.

Na-nun cheolsu-lul joha-han-ta

I-CASE Chulsoo-CASE LIKE-AUX-SENTEND

<SEED> anchor="좋아하-" nested-source="w-나" type="dir-explicit" subjectivity-type="emotion-pos" target="철수" polarity="pos" intensity="medium" </SEED>
Non-continuous Text Anchors

• Korean has a relatively free word order
• If “보다 좋다” ("like more") is taken as a single sentiment statement:

• A가 B보다 좋다
  • A-ka B-bota cotha
  • A-SUB B-COMPARATIVE good-SENTEND
  • A is better than B

• B보다 A가 좋다
  • B-boda A-ka cotha
  • B-COMPARATIVE A-SUB good-SENTEND
  • A is better than B
Hangul (한글)

- Phonetic alphabet where individual consonants and vowels are denoted by 자모 (jamo)
- *Jamo* combine into blocks representing a complete syllable
  - Called 글자 (*kulja*, “written character”) or 음절 (*umjeol*, “syllable”)

<table>
<thead>
<tr>
<th>Jamo</th>
<th>Syllable/Kulja</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ㄱ/</td>
<td>/ㅏ/</td>
</tr>
<tr>
<td>/h/</td>
<td>/a/</td>
</tr>
</tbody>
</table>
Complex Verbal Inflections

• 좋아다 (jota)
  • good-SENTEND [It’s] good.

• 좋아하다 (johahata)
  • good-AUX-SENTEND [I] like [it].

• 좋아졌다 (johajyeotta)
  • good-PROCESS-PAST-SENTEND [It] became good.

• 좋겠다 (jokhetta)
  • good-CONJECTURAL-SENTEND [That] will be good.

• For KSC, 좋 (joh) or 좋아 (joha, both “good”) is the important part. The inflection is irrelevant.
Inflecting Korean Verbs

- Inflections may cross syllable boundaries:

<table>
<thead>
<tr>
<th>하</th>
<th>+</th>
<th>-ㅂ니다</th>
<th>=</th>
<th>합니다</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ha</em></td>
<td><em>+</em></td>
<td><em>-mnida</em></td>
<td><em>=</em></td>
<td><em>hamnida</em></td>
</tr>
<tr>
<td>To do</td>
<td>(base form)</td>
<td>Sentence ending (formal polite)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Inflecting Korean Verbs

- Inflections may cause phonologic and orthographic reductions/assimilations:

<table>
<thead>
<tr>
<th>하</th>
<th>+</th>
<th>-았</th>
<th>+</th>
<th>-다</th>
<th>=</th>
<th>했다</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ha</em></td>
<td></td>
<td><em>-at</em></td>
<td></td>
<td><em>-ta</em></td>
<td></td>
<td><em>haetta</em></td>
</tr>
</tbody>
</table>
| To do (base form) | | Past tense particle | | Sentence ending (written neutral) | | }
Hangul as Unicode

- Unicode’s Hangul Syllables block contains 11172 *kulja*
  - U+AC00 through U+D7AF
  - Different Unicode blocks exist for individual *jamo*
- To a computer 하 (ha, U+D558) and 한 (han, U+D55C) are as different as A (U+0041) and B (U+0042)
- Unicode Hangul Syllable characters are atomic
  - A formula *does* exist to calculate constituent *jamo but* it is unreasonable to expect projects using the KSC to implement this themselves
Inflections and Annotations

- Inflection may not be relevant to annotation or inflection itself may be the only subjectivity expression
  - Tagging entire word may make automated analysis difficult
- Inflections in Korean can alter base form’s orthography; cannot expect users to solve this problem themselves

- Solution: Represent text using base morphemes
“Additionally, it was revealed that the imported Japanese Massushita Juicer’s power cable was 98cm, shorter than the standard 140cm.”
Base Morpheme Readability

- 수입품인 → 수입품 이 ∴
  - suimphum-i-n
  - imports-COPULA-ADJ

- 나타났다고 → 나타나 았 다고
  - nathana-at-tako
  - appear-PAST-REPORTING

- 밝혔다 → 밝히 었 다
  - palkhy-et-ta
  - reveal-PAST-SENTEND

- Difficult to read in large amounts
Annotation Requirements

- Ability to switch between fully-inflected text and base morphemes
  - Effectively doubles file size
- Ability to select non-continuous anchors
- Subjectivity Expression, Nested Sources, and Targets should all be created at the same time
  - Separating these attributes incorrectly implies they can exist independently
  - Also requires a counter-intuitive order of annotations
Existing Tools
GATE

• Available from http://gate.ac.uk
• Supports full range of Text Engineering Processes
  • From Annotation to Interpretation
• User-specifiable annotation schema
• Includes GATE Teamware
  • A web-based parallel annotation solution
GATE
brat

- Available from http://brat.nlpllab.org
- Web-based toolkit for annotation and visualization
2012.01.06 01:49  
버락 오바마 미국 대통령은 5일(현지시간) 새로운 국방전략을 바탕으로 한 군 개혁으로 미군을 보다 숭고화하지만 세계 최강군으로서의 위치를 유지할 것이라고 강조했다.

오바마 대통령은 어날 펜타곤에서 가진 국방부의 새 국방전략 발표식에 직접 참석, '미국 군대는 군사를 보다 원활히 나막강한 것이라며 하지만 미국은 모든 종류의 긴급사태와 위협에 준비대세를 갖추고, 날렵하고 유연한 군대를 가진 군사적 우위를 지속적으로 유지할 것'이라고 말했다.

버락 오바마 미국 대통령이 5일(현지시간) 펜타곤에서 가진 국방부의 새 국방전략 발표식에 참석해 연결하는 고 있다.

미국 역대 대통령이 펜타곤에서 회견을 갖고 연설을 한 것들은 극히 이례적으로 이번 새 국방전략이 오바마 대통령의 뜻에 따르 것이라는 점을 분명히 하려는 의지로 보인다.

AP연합뉴스
Drawbacks of Existing Tools

• Neither GATE nor brat allow for...
  • Switching between raw and parsed text
  • Annotating non-continuous segments of text
• In both Text Anchor, Nested Source, and Target must all be annotated separately then then...
  • Reference these separate annotation in a further “Sentiment Tag” annotation (GATE)
  • Create links between these separate annotations using a simple drag and drop technique (brat)
Design and Implementation
Design Motivation

• Structure of Korean Language and Orthography causes problems with many tools designed for Latin Alphabet Languages
• Lab undertook several text annotation-based projects
• Previous projects used custom quick-and-dirty tools that were purpose-built and difficult to adapt or maintain
Design Goals

• Focus on future adaptability and maintainability
• Modular design
  • Few inter-dependencies between components
• Publisher-Subscriber Pattern (PubSub)
  • Components don’t need to know where input will come to/where output goes to, only the shape of the data
Design Overview

```
Main Window

- DocumentPanel
- AnnotationPanel
- AnnotationAttrPanel
- DBConnection

  SentencePanel

  SentenceCtrl

  TokenCtrl
```
Design Overview

Main Window

SentencePanel

TokenCtrl

SentenceCtrl

DocumentPanel

AnnotationAttrPanel

AnnotationPanel
Token Selection w/o PubSub

User Selects a Token for Annotation

User -- Select --> TokenCtrl

TokenCtrl -- Update Selection --> SentencePanel

SentencePanel -- Update Selection --> DocumentPanel

DocumentPanel -- Update Selection --> MainWindow

MainWindow -- Update Selection --> AnnotationAttrPanel

www.websequencediagrams.com
Token Selection w/ PubSub

- MainWindow, DocumentPanel, and SentencePanel no longer involved in token selection
- Fewer dependencies
- Asynchronous
Implementation

- Standalone application easier to develop in short timeframe
- Cross platform technologies
  - Python 2.7
  - wxPython 2.9
- Emulates environment's native GUI API
- MORE?
Raw Text vs. Base Morpheme

Raw Text View

Morpheme View

영동포구 신길1동 '진짜구이' 주인 이미숙씨는 "시골에서 진짜구이를 먹어 본 어른들이 특히 좋아한다"고 말한다.

1022 때문에 자기가 좋아하는 주제를 통해 시간을 가지고 천천히 보는 것도 한 방법이다.

1427 처음 취직하였을 때 가족들의 반응은 '매우 기뻐했다'가 38.9%, '비교적 좋아했다'가 47%.

1485 최근 '내가 좋아하는 계절'로 영화계에 데뷔한 그는 어머니를 빼닮은 눈, 아버지와 비슷한 표정으로 관객을 매혹시키고 있다.

1714 몇 년 후 그 학생의 실력이 좋아서 악기를 바꿔야 할 때는 다시 선생이 그 악기를 사들이 D라든 새로운 학생에게 판다.

1826 Q2처음인 입술이 부르다고 몸살을 앓기도 했으나, 그 고비를 넘기고 나니가 오히려 건강이 좋아지고, 집의 아이들에게 '근면'을 가르칠 수 있다는 보람도 느낄 수 있었다.
Non-continuous Text Anchors
SEED Tag Attributes 1

1. Subjectivity Expression’s text anchor
2. Type of tag being made
3. How the subjectivity is expressed
4. Specific type of subjectivity being expressed
1. Source of the subjectivity (Imp means an implicit source. Out means generic source)

2. Object the subjectivity is directed towards

3. Annotators can mark annotations for review
Subjective Tag Attributes

Morphemes Selected

Tag
○ Seed ○ Subjective ○ Objective

Polarity
○ None ○ POS ○ NEG ○ NEUT ○ COMP

Intensity
○ None ○ Low ○ Medium ○ High

Comment

Confident

SAVE CLEAR DELETE
Objective Tag Attributes
Future Directions and Improvements
Data Independence

• No technical reason why the tool cannot be used with other morphologic rich languages as-is
• **BUT** tool currently requires raw and parsed text to be manually linked in database before annotation can begin
• The ability to populate database from user-specified sources would greatly increase usability
  • How? Must pay careful consideration to data’s expected structure
User Defined Schemas

• Currently uses hard-coded KSC annotation schema
• User Defined Schemas needed for project independence
• GATE uses W3C-compliant XML schemas
Collaborative Tools

• Currently must be run locally on a local database
• Manual synchronization becomes more impractical as number of users increases
• Web-based interfaces like GATE Teamware and brat more convenient for users
Increased Tool Support

• Support more database solutions
• Support a wider range of tasks, like GATE
  • Currently only supports Visualization and Annotation
References
