Building Korean Abstract Meaning Representation Corpus

Hyonsu Choe♬, Jiyoon Han, Hyejin Park, Teahwan Oh, Hansaem Kim

NCSOFT Corp. ♬, Yonsei University

At The Second International Workshop on Designing Meaning Representations (DMR 2020)

The first Korean AMR Corpus

- Annotated 1,253 sentences mostly newswire & Wikipedia texts via Korean Propbank
- Assisted Annotation: reduced manual search cost and omission than 'from scratch'
- We've reported Smatch 0.79 of Inter-annotation agreement and basic statistics.
- Major disagreement analysis on 4 different phenomena will be followed.

Abstract Meaning Representation (AMR) Corpora

- **English AMR**
  - *The Little Prince Corpus* (TLP) - providing the foundation for multilingual AMR Research
  - *Bio AMR Corpus* - known for proving its applicability in the biomedical domain
  - *AMR Annotation Release 3.0 (LDC2020T02)* - English AMR managed to enter a stable phase
- **Non-English AMR has begun its expansion**
  - *The Little Prince Corpus:* Chinese(Li et al., 2016), Brazil-Portuguese(Anchieta & Pardo, 2018)
  - General-purpose corpus: Chinese(cf. Wang et al., 2018 & Li et al., 2019) Brazil-Portuguese(Sobrevilla Cabezudo & Pardo, 2019)
  - Preliminary studies: Spanish(Migueles-Abraira et al., 2018), Vietnamese(Linh & Nguyen, 2019), Turkish(Azin & Eryigit, 2019)

Related works on Korean AMR

- **Preliminary study on several distinctive grammatical phenomena**(Choe et al., 2019a)
  - Copula construction & its negation, Case-stacking (multiple nominative & accusative)
  - Suggests guidelines to fill the gap between lexicalizations and regularizable meanings
- **Annotation Guidelines v1.0**(Choe et al., 2019b)
  - Localization of English AMR Specification v1.2.6
  - Reinforced Specific guidelines for Korean & representation examples (cf. honorifics, postpositions)

Korean AMR: Overview

- Broad adoption of special frames & entities of English AMR for intuitive annotation
  - Special frames & Entities: `have-org-role-91`, `include-91`, `rate-entity-91`, …
  - Reification & Constructions: `cause-01`, `exemplify-01`, `have-manner-91`, `correlate-01`, …
  - Modality: `obligate-01` for deontic modal construction such as "-야 하/되-"(have to (be)
  - Named-entity types: We took 'canonical NE types list' as a metalanguage of Korean AMR
    - Such as `person`, `country`, `government-organization`, …
    - And newly defined types like `brand`, `service`, `hospital`, `weapon`, …

Korean AMR: Overview

- Adjusted coverage of `:polarity -`
  - Some negator prefixes in derivative verbs are regularized in Korean AMR.
    - "불가능" (impossible) → `가능-01`(possible) `:polarity –`
    - "미완성" (incomplete) → `완성-01`(complete) `:polarity –`
  - 'Lexically negative verbs' are also regularized to represent polarity explicitly.
    - Antonym of copula '-이-': "X는 Y가 아니다" (X is not Y) → `Y :polarity – :domain X`
    - Antonym of '있-'(exist): "X가 없다" (X is not exist) → `X :polarity –`
    - Antonym of '알-'(know): "X를 모르다" (do not know X) → `알-01`(know) `:polarity - :ARG1 X`
    - Deontic modal negator verb '말-'(desist): "가지 마라." (Do not go.) → `가-01`(go) `:polarity - :mode imperative`

Korean AMR: Overview

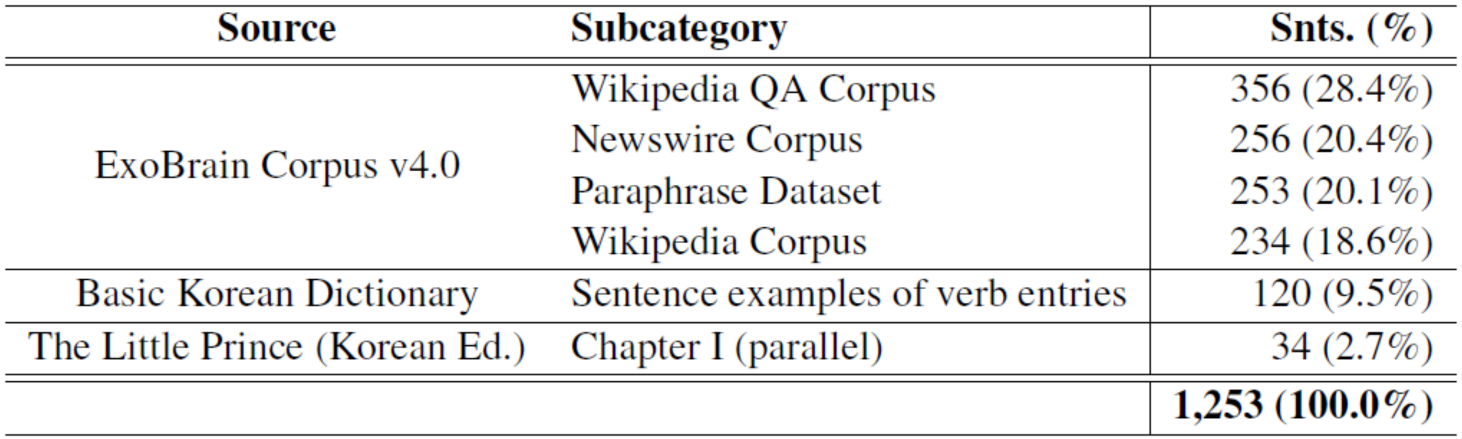
- Expanded use of `:domain` for case-stacking
  - Nominative/accusative marker can be licensed multiple times in a single sentence in Korean.
  - In most cases, a proper role can be assigned, but there are some outliers…
  - Multiple nominatives: 한국이 이웃들이 인심이 좋다. (In Korea, the neighbors are kind/generous.)
    *The yellows indicates nominative markers.

```
(x1 / 좋-01 # verb phrase
    :ARG1 (x2 / 인심)
    :domain (x3 / 이웃)  # plural marker '들' is dropped.
    :domain (x4 / country :name "한국")) # named-entity 'Korea'
```

  - Topical focus is annotated with `:domain` when it's hard to determine proper role from in-verb roleset.

Annotation dependencies

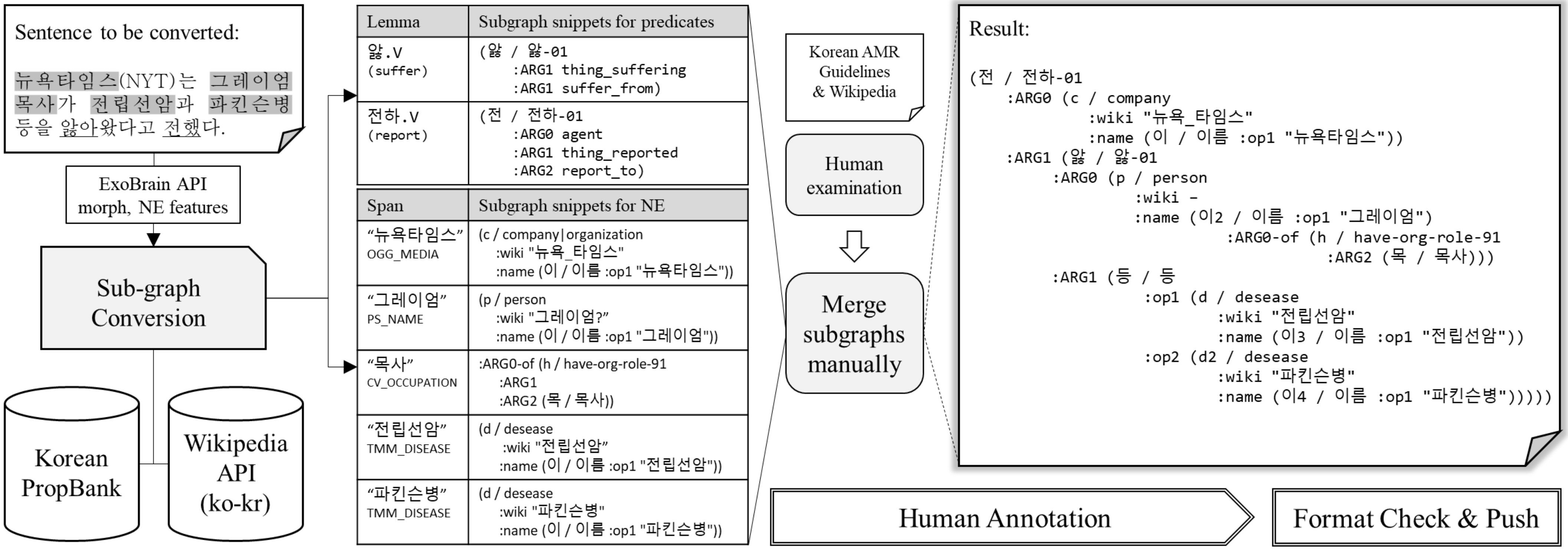
- Verb frames: Korean Propbank(Palmer et al., 2006)
  - 2,749 verb frames: covered 3,184 verb & adjectives (79.7% for tokens, 65.2% for types) in Korean AMR
  - If a verb frame is unavailable, we simply put '-00' and took the roleset of synonym verb frame.
- Source texts: (mainly) ETRI ExoBrain Corpus v4.0



Source	Subcategory	Snts. (%)
ExoBrain Corpus v4.0	Wikipedia QA Corpus	356 (28.4%)
	Newswire Corpus	256 (20.4%)
	Paraphrase Dataset	253 (20.1%)
	Wikipedia Corpus	234 (18.6%)
Basic Korean Dictionary	Sentence examples of verb entries	120 (9.5%)
The Little Prince (Korean Ed.)	Chapter I (parallel)	34 (2.7%)
		1,253 (100.0%)

Assisted annotation

- Subgraph snippets are preprocessed from POS & NE features → Human annotator merged them into a solid AMR.
- The cost for searching Propbank and Wikipedia & unintentional omission had reduced apparently.



Sentence to be converted:

뉴욕타임스(NYT)는 그레이엄 목사가 전립선암과 파킨슨병 등을 앓아왔다고 전했다.

Lemma	Subgraph snippets for predicates
앓.V (suffer)	(앓 / 앓-01 :ARG1 thing_suffering :ARG1 suffer_from)
전하.V (report)	(전 / 전하-01 :ARG0 agent :ARG1 thing_reported :ARG2 report_to)

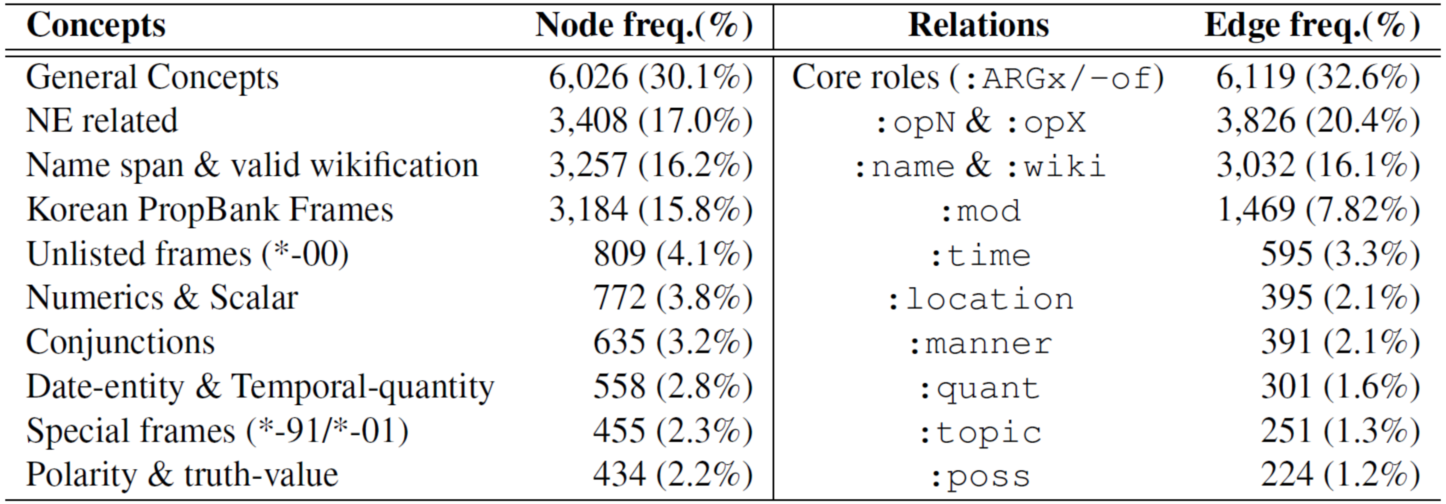
Span	Subgraph snippets for NE
"뉴욕타임스" OGG_MEDIA	(c / company|organization :wiki "뉴욕_타임스" :name (이 / 이름 :op1 "뉴욕타임스"))
"그레이엄" PS_NAME	(p / person :wiki "그레이엄?" :name (이 / 이름 :op1 "그레이엄"))
"목사" CV_OCCUPATION	:ARG0-of (h / have-org-role-91 :ARG1 :ARG2 (목 / 목사))
"전립선암" TMM_DISEASE	(d / desease :wiki "전립선암" :name (이 / 이름 :op1 "전립선암"))
"파킨슨병" TMM_DISEASE	(d / desease :wiki "파킨슨병" :name (이 / 이름 :op1 "파킨슨병"))

Result:

```
(전 / 전하-01
    :ARG0 (c / company
             :wiki "뉴욕_타임스"
             :name (이 / 이름 :op1 "뉴욕타임스"))
    :ARG1 (앓 / 앓-01
         :ARG0 (p / person
                  :wiki -
                  :name (이2 / 이름 :op1 "그레이엄")
                            :ARG0-of (h / have-org-role-91
                                         :ARG2 (목 / 목사)))
         :ARG1 (등 / 등
                  :op1 (d / desease
                          :wiki "전립선암"
                          :name (이3 / 이름 :op1 "전립선암"))
                  :op2 (d2 / desease
                          :wiki "파킨슨병"
                          :name (이4 / 이름 :op1 "파킨슨병")))))
```

Korean AMR Corpus

Statistics



Concepts	Node freq.(%)	Relations	Edge freq.(%)
General Concepts	6,026 (30.1%)	Core roles (`:ARGx/-of`)	6,119 (32.6%)
NE related	3,408 (17.0%)	`:opN` & `:opX`	3,826 (20.4%)
Name span & valid wikification	3,257 (16.2%)	`:name` & `:wiki`	3,032 (16.1%)
Korean PropBank Frames	3,184 (15.8%)	`:mod`	1,469 (7.82%)
Unlisted frames (*-00)	809 (4.1%)	`:time`	595 (3.3%)
Numerics & Scalar	772 (3.8%)	`:location`	395 (2.1%)
Conjunctions	635 (3.2%)	`:manner`	391 (2.1%)
Date-entity & Temporal-quantity	558 (2.8%)	`:quant`	301 (1.6%)
Special frames (*-91/*-01)	455 (2.3%)	`:topic`	251 (1.3%)
Polarity & truth-value	434 (2.2%)	`:poss`	224 (1.2%)

- NE-related concepts account for 17.0%, including newly defined NE types: `brand`, `service`, `hospital`, `weapon`, …
- 3,184 verb & adjective instances (15.8%) are valid Korean Propbank frames, while 809 instances (4.1%) are not.
- The core-roles (and its inverse roles) account for 32.6% in order of `:ARG1`, `:ARG0`, `:ARG2`, `:ARG3`, …
- The Inter-annotator Agreement (IAA) of 4 annotators, based on 50 sentences, reached Smatch 0.79*

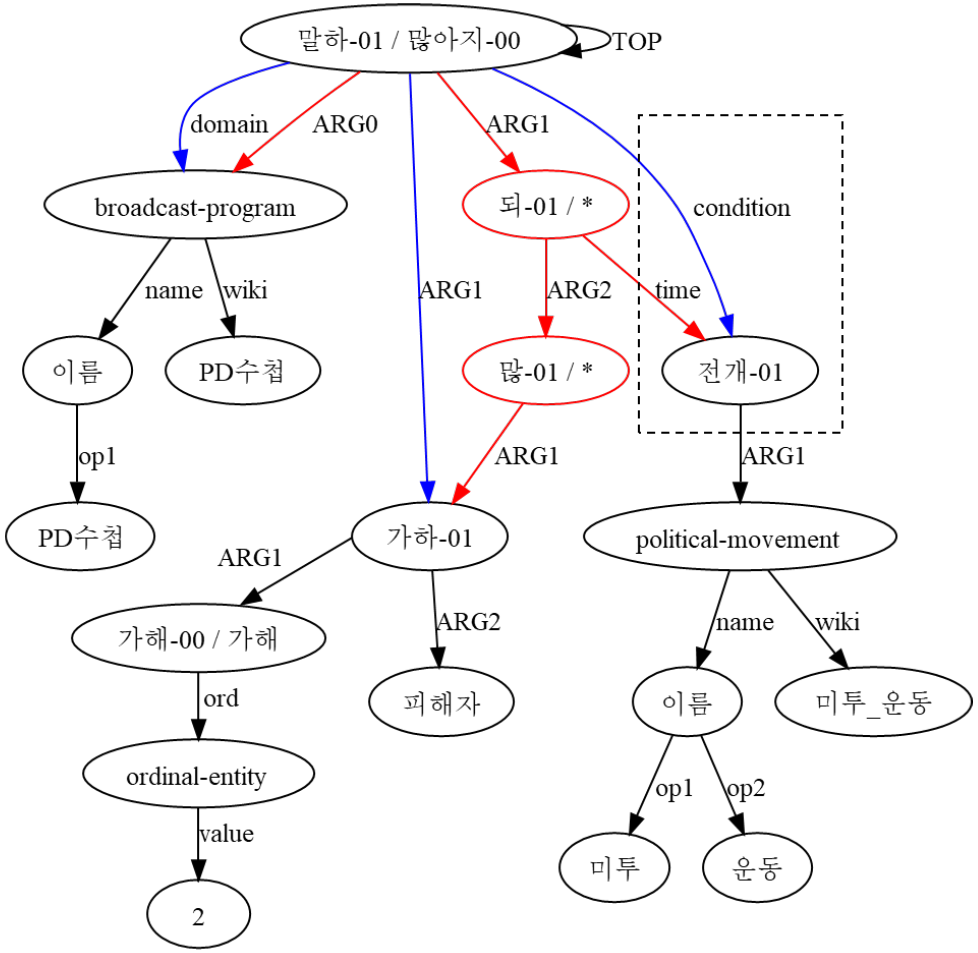
* In English: 0.79 ~ 0.83
In Brazil-Portuguese: 0.72

Disagreement Analysis

Annotators showed conflicting views on some representations

- Adverbial clause marker
- Conjunctional suffix
- Special postposition (non-case-marking particles)
- Collocation (N+V complex predicate)

Case 1. Interpretation of adverbial clause marker



"미투 운동이 전개되면서, 피해자들에게 2차 가해를 가하는 것도 많아지고 있다."

"As Me-Too movement expands, more and more victim-blamings are inflicted to the victims."

- Discrepancy on representing relation between preceding and following clauses.

- The red simply interpreted the clause marker '-면서' to the relation of `:time`.

- Otherwise the blue interpreted the clause marker to the relation of `:condition`, in the sense of *'according as'* or vague correlation.

Case 2. Conjunctional suffix (coordination vs. subordination)

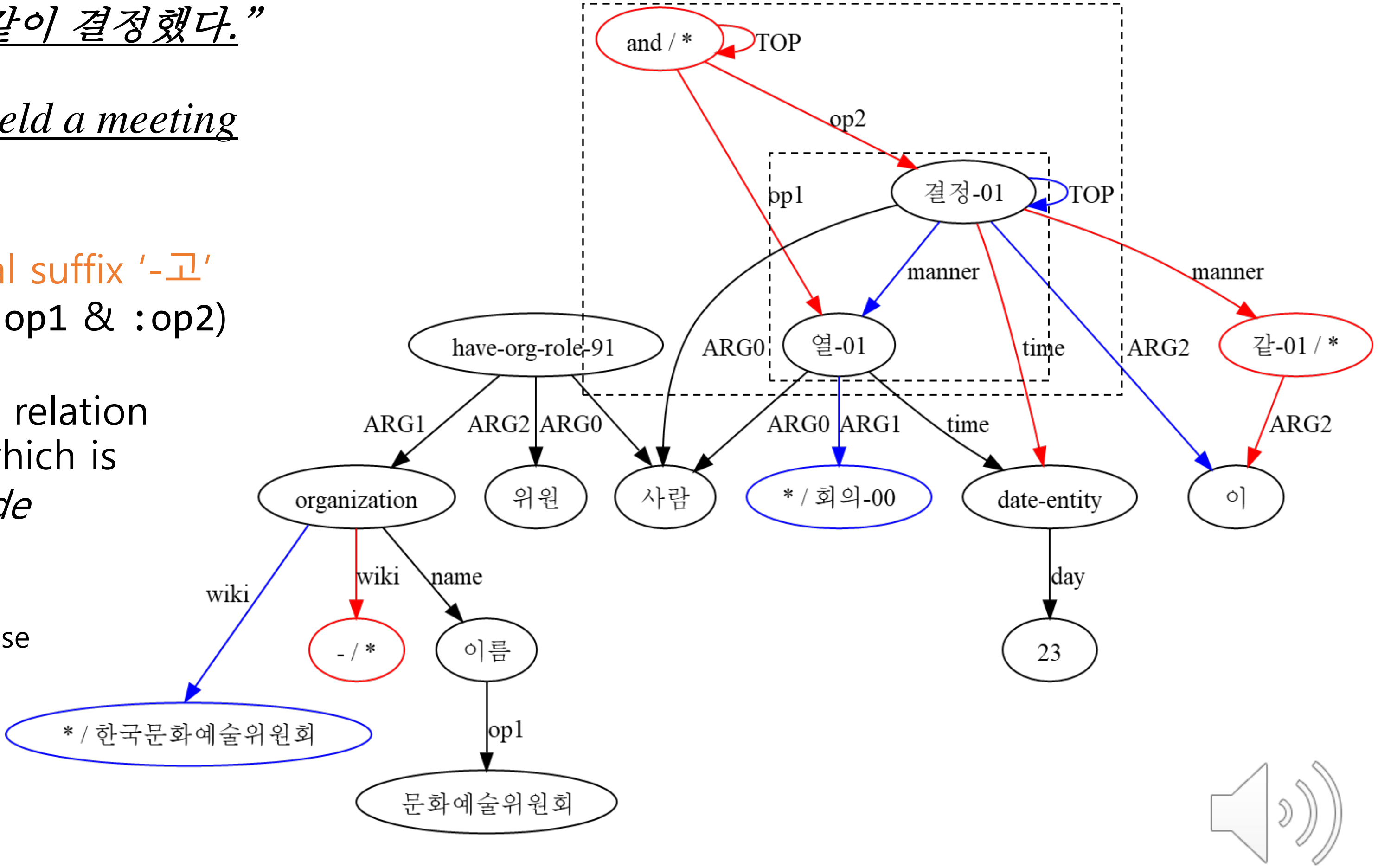
"문화예술위원회 위원들은 23일 회의를 열고 이같이 결정했다."

"The members of the Culture and Arts Council held a meeting on the 23rd and made the decision."

- The red simply represented the conjunctional suffix '-고' as the root (and) of two conjoined events. (`:op1` & `:op2`)

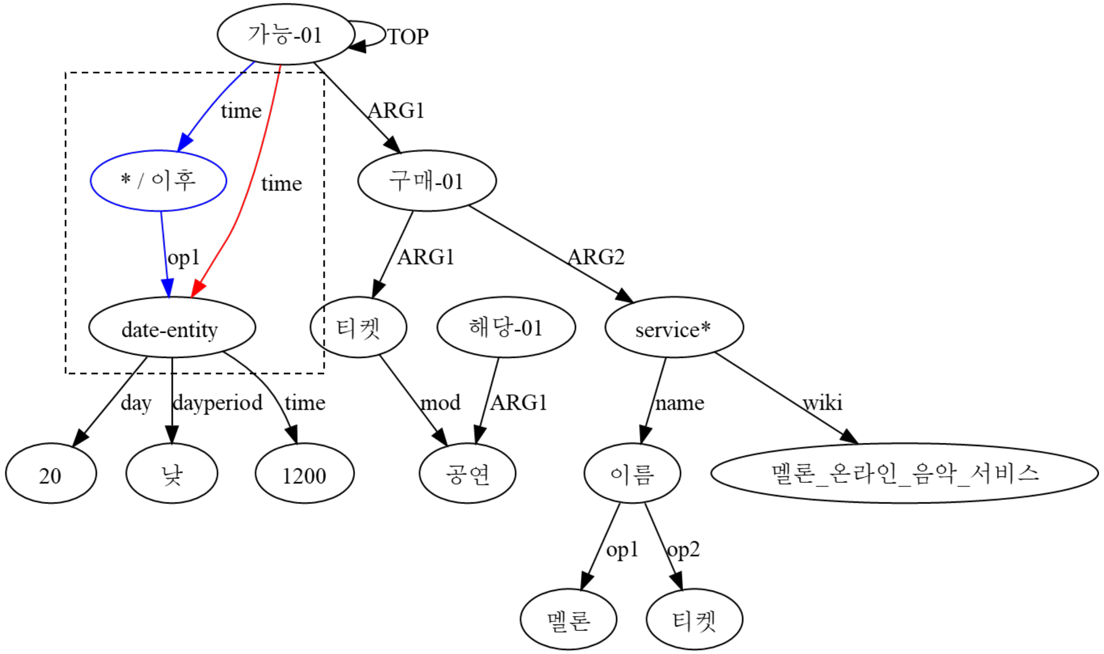
- The blue represented the conjunction as the relation of `:manner`, in the sense of subordination, which is can be interpreted as *'The decision was made in the manner of holding the meeting.'*

- If past tense markers are realized on the verbs of each clause like *"회의를 열었고 이같이 결정했다."*, it can be interpreted as 'two events had occurred separately.'



Case 3. Special postposition

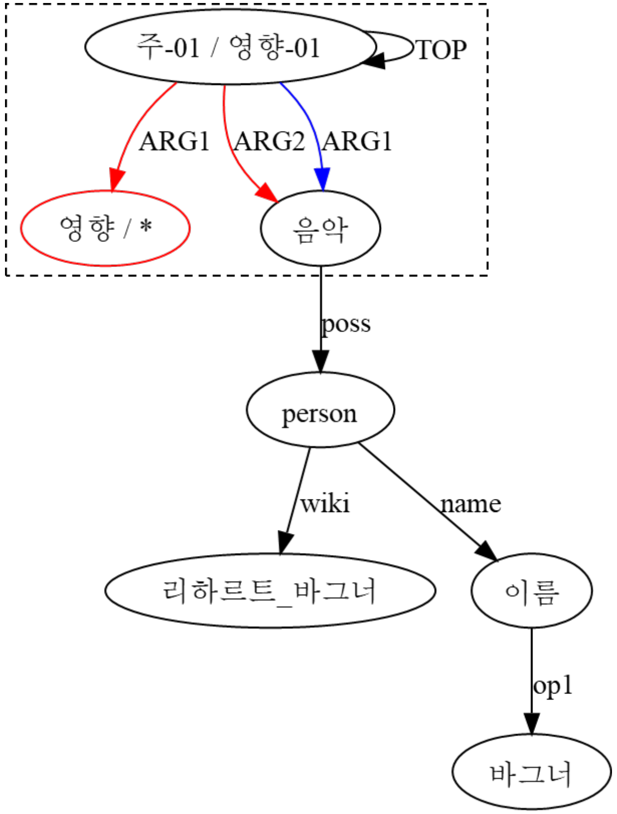
- There are over 30 types of special postposition which are marking non-core role and add nuances.
- The red simply represented postposition '-부터' as a relation `:time`, in the mere sense of the time of the event.
- The blue represented the postposition with the interpolation 이후(after), in the sense of the starting point that invokes a specific time span.
- It is not easy to decide when the nuance can be abstracted away or not.
- Proposing representation for every use of specific post-positions requires broad investigation like English AMR did.



“해당 공연 티켓은 멜론 티켓에서 20일 낮 12시부터 구매할 수 있다.”

“The tickets for the performance can be purchased from 12 p.m. on the 20th at Melon Ticket”

Case 4. Collocation (N+V complex predicate)



"바그너의 음악에 영향을 주었다."

"It influenced Wagner's music."

- Did missing lexical counterpart in semantic opposite lead a discrepancies?

	Derivative verb lemma	Deverbal noun structure
Subj → patient (influenced)	영향받-.v *be influenced*	영향을 받-. vp *get influence*
Subj → agent (influencer)	- *influence*	영향을 주-.vp *give influence*

받-take

주-give

영향-01$^{\text{influence}}$?

주-01$^{\text{give}}$ + 영향$^{\text{influence}}$?

- The red represented the N+V complex predicate following its deverbal noun structure: noun concept 영향$^{\text{influence.n}}$ + main verb 주-01$^{\text{give.v}}$.
- The blue simply represented the N+V complex predicate with 영향-01$^{\text{influence.v}}$ despite of it has the only lexical mapping to *영향받-.v* in Korean Propbank.

Final remarks

- Lack of MWE predicate frames, Lack of case-by-case guidelines for picky cases are big issue.
- Limited size of the corpus is also needed to be resolved for further applications.
- Further works should be followed for the quality and quantity of Korean AMR.
- The corpus is available at: https://github.com/choe-hyonsu-gabrielle/korean-amr-corpus
- Please send me an e-mail: choehyonsu@gmail.com, I'll reply as fast as I can!