KoCoNovel 데이터셋
상호참조해결 가이드라인(ver.0.1)

작성자: 김규희

1. 전처리(Preprocessing)

전처리 작업으로 텍스트 내부의 오타와 잘못된 줄 구분을 수정했습니다. 더불어 데이터셋의 범용성과 주석 작업자들의 편의를 고려해 현대 한국어 문법에 맞게 맞춤법을 교정하는 작업을 진행했습니다. 맞춤법 교정과정에서는 부산대 한국어 문법 검사기를 활용했고, 연구팀이 수동으로 검수하였습니다. 작업자의 편의를 위해 20문장 내외로 장면을 나눠, 주석작업을 진행했습니다.

2. 참고자료(Guideline References)

KoCoNovel 데이터셋의 가이드라인은 ETRI 상호참조해결 가이드라인과 국립국어원 모두의 말뭉치 상호참조해결 말뭉치의 가이드라인을 기초로 제작되었습니다. **따라서 본 가이드라인에서 제시되지 않은 항목에 대해서는 ETRI나 국립국어원의 가이드라인을 참고하시길 바랍니다.**

3. 주석 대상(Mention)

KoCoNovel 데이터셋에서는 명사, 대명사, 명사구로 된, 소설 속 등장인물을 지시하는 멘션을 주석했습니다. 멘션이 될 수 있는 대상, 그리고 멘션의 범위인 스팬(Span)에 대한 가이드라인은 아래와 같습니다.

3.1. 등장인물(Definition of Character)

등장인물은 일반적으로 인간이지만, 어떤 경우에는 동물이거나 심지어 다른 것일 수도 있습니다. KoCoNovel 데이터셋에서는 다음과 같이 등장인물을 정의했습니다.

- **사람인 경우(Human)**
  멘션이 지칭하는 개체가 특정한 사람인 경우, 등장인물 멘션으로 취급했습니다. 하지만, 일반적으로 등장인물은 'A person who appears in a story, book, play, movie, or television show'(Britannica)으로 정의되기에 'Appearance'의 측면에서, 대화 상에서 한 번만 언급되는 대상을 등장인물로 볼 수 있는지에 대해 논란의 여지가 있습니다. 그럼에도 액자식 구성을 고려할 때, 'Appearance'를 기준으로 등장인물을 정의할 수 없다고 판단, 다음의 경우를 제외하고는 모든 인물 관련 멘션에 대해 주석하도록 하였습니다.

- 공자, 맹자, 예수 등 학문 및 종교와 관련해, 관용 표현으로 쓰이는 경우
- 고구려 대무신왕 십오년' 등 연도를 나타내는 관용 표현으로 쓰이는 경우

또한 사람이 모여 만들어진 집단에 대해서는 독립된 법인격으로 취급되는지를 기준으로 판단했습니다. 예를 들어, 별도의 법인격을 가지는 '잡지 본사'와 같은 회사는 등장인물 멘션에 해당하지 않지만, '군중', '소년 무리' 등의 집단은 모두 등장인물 멘션에 속한다고 보았습니다.

- **사람이 아닌 경우(Non-human)**
  멘션이 지칭하는 개체가 사람이 아닌 경우, 우리는 대사를 하는지, 혹은 다른 등장인물과 감정적 상호작용을 하는지를 기준으로 등장인물 여부를 판단하였습니다. 예시는 다음과 같습니다.

예시	등장인물 취급 여부	판단 근거
사람에게 말을 걸어 부탁하는 도깨비	O	대사가 존재함.
(말을 하지 않지만) 주인공의 말에 고개를 끄덕이고, 눈물을 흘리는 소	O	주인공과 감정적 교류를 하기 때문.
주인공이 자신을 위해 울어준다고 생각하는 새	X	주인공의 감정이 투영된 것에 불과함.
주인공이 귀신이라고 착각해서 무서워하는, (대사와 행동 묘사가 존재하지 않는) 대상	△	실제로 지시하는 개체가 무엇인지에 따라 달라짐. 주인공이 일방적으로 무서워하는 것은 감정교류에 해당하지 않음.

- **독자(Reader)**
  일부 소설에서는 서술자가 직접적으로 독자를 부르는 경우가 존재합니다. 이에 등장인물은 아니지만, 별도의 레이블 <READER>를 사용하여 멘션과 상호참조해결을 주석하였습니다.

  (e.g.) *현명한 **독자**여! 무엇을 주저하는가. 이 중하고도 큰 문제는 **독자**의 자각과 지혜와 힘을 기다리고 있지 않은가!* (이효석, '도시와 유령')

3.2. 스팬의 단위(Span)

KoCoNovel 데이터셋에서는 주석 범위인 스팬(Span)의 단위를 어절이 아닌 형태 단위로 하였습니다. ETRI 가이드라인에서는 어절 단위로 멘션을 표시하였으나, 국립국어원의 가이드라인에 따르면, 이는 주석자의 직관과 일치하지 않을 뿐더러 불필요한 조사를 포함하게 된다는 문제가 있습니다. 따라서 KoCoNovel 데이터셋에서는 국립국어원의 가이드라인에 따라

형태 단위로 맨션을 표기했습니다. 단, 음절 이하의 단위에서 형태소를 분리해서 주석하지는 않았습니다. (e.g. '날'('나' + '-ㄹ' → '날'), '우린'('우리' + '-ㄴ' → '우린'))

3.3. 최대 스팬 원칙(Maximal Span)

다음으로, 최대 스팬(Maximal Span) 원칙을 따르되, 예외 케이스를 수정하였습니다. 최대 스팬 원칙은 모든 수식어구를 포함해 최대의 범위로 멘션을 표시하는 것으로, 대부분의 상호참조해결 데이터셋에서 준수되고 있는 규칙입니다. ETRI와 국립국어원 모두 최대 스팬의 원칙을 따르고 있지만, 다음과 같은 경우에 대해서는 예외 조건을 설정하였습니다.

(1) [이름+직함]이 결합한 경우, 내포된 이름을 멘션으로 한 번 더 추출한다.
(2) 한정사구/지시 형용사*에 의해 수식을 받거나 대명사인 경우, 이전의 수식어구를 스팬에 포함하지 않는다.

KoCoNovel 데이터셋에서는 (1)의 조건이 인명이 많은 데이터셋의 특성상 주석과정에서 많은 불편함을 야기하고 무의미한 상호참조관계를 증식시킨다고 판단했습니다. 따라서 (1)의 조건을 예외 조건에서 삭제하여 내포된 이름을 멘션으로 한번 더 추출하지 않고, 최대 스팬만을 표시하였습니다. 또한 (2)의 조건에, 관형사절에 의해 고유명사, 혹은 고유명사처럼 사용되는 명사구가 수식받는 경우**에는 관형사절을 스팬에 포함하지 않도록 했습니다. 이는 한국어의 통사적 특징을 고려한 것이자, 스팬이 15 토큰 이상인 멘션을 최소화하기 위한 것입니다.

*영어에서는 한정사구(Determiner Phrase)가 명사구의 범위를 쉽게 한정짓는 반면에, 한국어에서는 관사(article)를 포함해 한정사구라고 할 수 있는 것이 존재하지 않는다. ETRI는 영어 가이드라인을 번역에 도입하는 과정에서 DP를 직역하여 한정사구라고 하였고, 국립국어원에서는 이를 수정, 지시형용사를 기준으로 Maximal Span을 결정했다.

*고유명사처럼 사용되는 명사구의 판단 기준은 §3.5 참고

3.4. 싱글톤(Singleton)

싱글톤은 상호참조관계를 가지지 않는 멘션으로, 이것의 포함여부는 데이터셋마다 다릅니다. KoCoNovel 데이터셋에서는 **상호참조해결 외에도 등장인물 개체명 인식(NER)을 수행할 수 있도록, 싱글톤 또한 모두 포함했습니다.**

3.5. 일반 멘션과 특정 멘션의 구분(Generic vs Specific)

명사구는 때로 특정한 개체가 아닌 일반적인 클래스를 가리킬 수 있으며, 이러한 명사구를 일반멘션(Generic mention)이라고 합니다. OntoNotes(Hovy et al., 2006)에 따르면, 영어에서 일반 멘션은 관사가 붙지 않은 복수형 혹은 부정관사로 시작되는 부정 명사구입니다. 한편, 한국어에서는 관사를 비롯해 한정사구에 해당하는 것이 존재하지 않고, 고유명사에 대해

특정한 표기상의 구분이 없으며, 단수와 복수를 표현하는 수 표현이 발달하지 않아 일반멘션인지 특정멘션인지 구분이 어렵습니다. 또한 한국어에서는 이름보다는 사회적 관계에 따른 다양한 방식의 호칭어(e.g. 어머니)를 사용하는 것을 선호하는 호칭문화가 존재하여, 문맥에 따라 이 둘을 구분해야 합니다. 국립국어원에서는 일반 멘션과 특정 멘션에 대한 판단을 돕기 위해, 다음과 같은 기준들을 제시했습니다.

- 이어진 문장들에서 동일한 어구의 형태로 계속 반복적으로 나타나는가?
- 이어진 문장들에서 줄임말, 유의어 등을 사용한 동일 대상 지칭 변용 표현들이 나타나는가?
- 텍스트의 내용과 밀접히 관련된 핵심적인 명사구인가?

KoCoNovel 데이터셋에서는 이에 더해 다음과 같은 규칙들을 정했습니다.

1. 일반멘션은 구체적인 지칭대상이 없기에 주석대상에서 제외하지만, '독자'는 예외로 한다.

2. 수를 알 수 없더라도 맥락에 의해, 시간과 장소가 특정될 수 있는 집단(e.g. 광장에 모인 사람들, 주인공의 땅을 파고 있는 농민들)은 특정 멘션으로 취급한다.

3. 아무런 한정 수식이 없는 일반 명사도 소설 내에서 특정 개체를 반복적으로 지칭한다고 판단되면 특정 멘션으로 취급한다. 그리고 소설 속에서 다른 개체를 지칭할 가능성 없을 경우, 고유명사에 준하는 것으로 취급하고 수식어를 제외하였다.

   (예시1) 김동인의 소설 <대동강은 속삭인다>에서 주인공의 어머니는 서술자에 의해 아무런 수식 없이 '어머니'로만 표현되고, '어머니'라는 단어는 다른 개체를 지칭하는데 사용되지 않는다. 이 경우, '어머니'는 고유명사이자, 특정 멘션으로 취급한다.

   (예시2) 이광수의 소설 <모르는 여인>에서도 아무런 한정 수식 없이 '여동생'이라는 단어가 주인공의 여동생을 지시하기 위해 사용되었다. 그러나 주인공에게는 이전에 죽은 다른 여동생이 한 명 더 존재했다. 따라서 맥락에 따라 특정멘션으로 취급하되, 고유명사로는 취급하지 않는다.

4. '세상에 둘도 없는'과 같은 표현이나 지시형용사가 쓰인 명사구는 특정멘션으로 취급한다.

3.6. 고유 명사구 내부 멘션(Nested Mentions in Proper Noun Phrases)

OntoNotes(Hovy et al., 2006)에 따르면, 더 이상 나눠질 수 없다고 판단되는 고유명사구의 경우, 그 내부에 다른 개체를 가리키는 멘션이 있더라도 표시하지 않도록 합니다.(e.g. 'Bank of America') 한편, 한국 문화에서는 '보부 엄마'와 같이, 자식이 있는 여성을 '그녀의 자식 이름(주로 첫째 아이)+엄마'로 부르는 경향이 존재합니다. 이는 많은 사회적 관계에서 여성의 이름을 완전히 대신하며, 고정적인 형태로 특정한 단일 대상을 지칭한다는 점에서 고유명사라고 볼 수 있습니다. KoCoNovel 데이터셋에서는 해당 호칭을 고유 명사구로 처리하되, 내부에 위치한 자식 이름에 대해서도 주석을 허용했습니다.

3.7. 소유격 대명사(Possessive)

기본적으로 소유격 대명사는 대명사의 한 종류로 멘션에 포함됩니다. 그러나 한국어에서 '우리'와 '저희'는 지칭 대상이 모호한 경우가 존재합니다. ETRI와 국립국어원의 가이드라인에 따르면, '우리'와 '저희'는 일반 멘션으로 사용될 경우 주석대상에서 제외됩니다. 이에 더해, KoCoNovel 데이터셋에서는 '우리'가 특정멘션이더라도 모두 제외하였습니다. 이들 표현이 내집단과 외집단을 구분할 때 사용되기에, 등장인물 관계에 대한 주석자의 주관적 판단이 개입될 여지가 많다고 보았기 때문입니다. 예를 들어, 시어머니와 며느리의 대화 중, 며느리의 발화에서 '우리 승호(아들 이름)'라는 표현이 나왔을 때, '우리'가 누구인지는 둘의 관계에 대한 판단에 따라 달라집니다. 따라서 KoCoNovel 데이터셋에서는 이러한 '우리'에 대해서는 주석하지 않고 모두 제외하였습니다.

4. 상호참조관계(Coreference)

상호참조관계는 기본적으로 동일 지시체(Coreferent)를 갖는 멘션들을 찾는 것을 의미합니다. 상호참조관계의 특수한 경우에 대한 규칙들은 다음과 같습니다.

4.1. 코퓰래와 동격(Copulae and Apposition)

코퓰래와 동격을 상호참조관계로 볼 것인지에 대한 판단은 데이터셋마다 다릅니다. 코퓰래와 동격을 아예 주석하지 않기도, 'COP', 'APPOS'와 같은 특수한 레이블을 사용해 주석하기도 합니다. ETRI와 국립국어원이 지정사(VNP)가 포함된 'A is B'(A는 B이다)라는 문장에 대해서 무조건적으로 수식어를 포함한 멘션 B를 추출하도록 한 것과 달리, KoCoNovel 데이터셋에서는 구문 구조와는 관계없이 특정 멘션이라고 볼 수 있는 것에 대해서만 상호참조관계를 인정하였습니다.

4.2. 단체와 개인의 상호참조해결(Coreference between Groups and Individuals)

많은 상호참조해결 데이터셋은 복수의 개체(e.g. '세 처녀')를 지칭하는 멘션을 단수 개체(e.g. '첫째 처녀', '둘째 처녀', '셋째 처녀')를 지칭하는 멘션과 독립된 것으로 간주하고, 서로 분리하여 주석했습니다. 그러나 문학 텍스트의 내용을 분석하기 위해서는 포함관계가 존재하는 두 멘션에 대해 그 관계를 주석할 필요성이 존재합니다.

이에 KoCoNovel 데이터셋에서는 선행 연구 모델들을 활용할 수 있도록 (1) 둘을 분리해 주석한 버전을 제공하는 한편, (2) 복수 개체와 단수 개체를 겹쳐 주석한 버전 또한 제공합니다. 특히 '어머니와 아들'은 '모자', '아버지와 아들'은 '부자', '어머니와 딸'은 '모녀', '아버지와 딸'은 '부녀'로 한 단어로 표현하는 경향이 있는 한국어의 어휘 특성을 고려할 때, (2)는 (1)에서는 드러나지 않는 많은 상호참조관계를 보여줍니다. 단, (2)에서도 단체를 이루는 개인이 명확하지 않거나, 단체 자체가 독립된 정체성을 지니는 경우에는 상호참조관계를 주석하지 않았습니다.

문장	분리해서 주석(1)	겹쳐서 주석(2)
나는 승호를 등에 업었다. 우리 모자는 눈 내리는 거리를 방황했다.	[‘나’], [‘승호’] [‘우리 모자’]	[‘나’, ‘우리 모자’], [‘승호’, ‘우리 모자’]
나는 독립군에 합류했다. 우리는 산으로 도망다녔다.	[‘나’], [‘우리’]	[‘나’, ‘우리’]

5. 문학 작품 특수성(Special Features of Literary Text)

문학 텍스트는 개체의 변화 및 정보의 비대칭성과 관련해 별도의 상호참조 규칙이 필요합니다. 이에 대한 내용은 아래와 같습니다.

5.1. 개체의 변화(Changes in Entity)

전통적인 상호참조해결 데이터셋에서는 두 개의 멘션이 실제세계(Real World)에서 동일한 개체를 지시하는지를 기준으로 판단하지만, 문학 텍스트에서는 멘션이 지시하는 대상을 실제세계에 대응시킬 수 없거나, 긴 서술시간으로 인해 정체성이 변화, 혹은 동일 개체임에도 서술자에 의해 변화와 차이가 강조되어 묘사되는 경우가 존재합니다. 이에 문학작품을 기반으로 하는 상호참조해결 데이터셋에서는 실제 세계와 관계없이 작품 세계(Text World)를 기준으로 개체의 동일여부를 판단합니다. 다만, 서술자의 묘사에 따른 정체성 변화 여부에 대해서는 서로 다른 기준을 가지는데, KoCoNovel 데이터셋에서는 서술자의 묘사에 따른 정체성 변화는 감안하지 않고, 작품 세계 내 동일 개체, 동일 정체성을 기준으로 상호참조관계를 주석했습니다.

> (e.g.) *정희는 하숙인 교회로 향하는 쓸쓸한 전차에 앉아서 **A**의 생각을 하면서, **그 얌전하고 사기 없고 쾌활하던 사랑스러운 계집애**를 **이런 비속된 여인**으로 변케 한 '시대'라는 것을 밉게 여겼다.*(김동인, ‘정희’)

5.2. 정보의 비대칭성(Asymmetry in Knowledge)

소설에서는 독자, 서술자, 캐릭터가 각각 가지고 있는 정보에 따라 다른 세계를 구축하기에, 어떤 시점을 채택하는가에 따라, 상호참조관계에 대한 주석은 달라집니다.기존의 문학작품 기반 상호참조해결 데이터셋은 전말을 알고 있는 전지적 작가 시점을 채택했습니다. 한편, KoCoNovel 데이터셋에서는 독자와 전지적 시점에서 데이터를 두 가지 버전으로 제공합니다. 독자 시점은 장면 단위로 정보가 제한된 반면, 전지적 시점은 작품 세계를 기준으로 주석작업이 진행되었습니다. 이는 장면 단위로 등장인물에 대한 정보 변화가 어떻게 이뤄지는지에 대해 보다 풍부한 정보를 제공하기 위함입니다.

KoCoNovel은 함께 만들어가는 데이터셋입니다. 가이드라인에 대한 수정 제안이 있다면 Github이나 이메일을 통해 연락주시길 바랍니다.