NerKor annotációs séma

Simon Eszter

1. Az annotáció formátuma

A korpusz végső formátuma a CoNLL-U Plus formátum lesz[1]. Ez a Universal Dependencies (UD)[2] projektben alkalmazott sztenderd CoNLL-U fájlformátum[3], de néhány dologban eltér tőle:

- Nullánál több, de amúgy bárhány oszlopa lehet.
- A CoNLL-U 10 előre definiált oszlopa közül bárhányat, akár nullát is tartalmazhat. Ezek mellett pedig új, projektspecifikus oszlopokat is tartalmazhat.
- A fájl első sorában jelölni kell, hogy milyen oszlopai vannak, például: `# global.columns = ID FORM LEMMA UPOS XPOS FEATS CONLL:NER`
- A CoNLL-U-s előre definiált oszlopok nevei változatlanok. Az új, projektspecifikus oszlopok nevében jelölni kell, hogy milyen névtérhez kapcsolódik az az annotáció, és hogy milyen típusú annotáció. A két információt egy kettőspont választja el egymástól, például: `CONLL:NER`.
- A fájlnevek kiterjesztése: `.conllup`.

A CoNLL-U Plus formátum végső soron egy `tsv` formátum, amiben egy sorban egy token szerepel, és a mondathatárt üres sor jelöli. A tabbal elválasztott oszlopokban pedig a fent meghatározott egyes annotációtípusok szerepelnek. Amennyiben egy annotációtípus az adott cellában kitöltetlen marad, akkor alulvonással ('_') kell jelölni.

A korpusz egy részén lesz morfológiai annotáció is. Ahhoz, hogy az egész korpusznak meglegyen az egységes formátuma, az egész korpuszban ugyanazokkal az oszlopokkal kell dolgozni. Így a morfológiai információt hordozó oszlopok a korpusz egy részében alulvonást fognak tartalmazni.

Az oszlopaink tehát: `FORM LEMMA UPOS XPOS FEATS CONLL:NER`, ahol

`FORM` maga a token;

`LEMMA` a token lemmája;

[1] `https://universaldependencies.org/ext-format.html`
[2] `https://universaldependencies.org/`
[3] `https://universaldependencies.org/format.html`

`UPOS` a UD jelölési formalizmusa szerinti szófajkód;

`XPOS` az `emMorph` [Novák et al., 2016] által kiadott morfológiai kód, amely tartalmazza a szófajkódot és a morfoszintaktikai információkat is;

`FEATS` a UD jelölési formalizmusa szerinti morfoszintaktikai jegyek;

`CONLL:NER` a NE annotáció.

2. NE annotáció

A fő annotációhalmazunk a named entity (NE). A 2002-es és 2003-as CoNLL shared taskok [Tjong Kim Sang, 2002, Tjong Kim Sang and De Meulder, 2003] sztenderd címkekészletét használjuk, vagyis 4 névkategóriát különítünk el: `PER`, `ORG`, `LOC`, `MISC`.

A CoNLL2002 annotációs formátumát, az ún. IOB2 formátumot követjük. Eszerint minden név első eleme '`B-`' prefixet, míg minden nem első elem '`I-`' prefixet kap. A nem neveket `O` betűvel jelöljük. Figyelem: ez egy nagy '`O`' betű, nem egy nulla! Ennek a kiértékelésnél van szerepe: minden címke `string` kell, hogy legyen, nem `integer`.

A NE annotációnak a teljes szöveget le kell fednie. Ez praktikusan azt jelenti, hogy minden tokenhez tartozó cellának ki kell lennie töltve, vagyis minden szövegelemet annotálni kell valahogy. Minden ugyanolyan értékű tokennek számít, az írásjelek is. A folyó szöveg tokenekre bontásakor bizonyos tapadó írásjeleket leválasztunk az előtte-utána álló szóról, így azok is önálló tokenné válnak. Minden, ami nem név, az kell, hogy kapjon `O` annotációt, így az írásjelek is. A mondathatárt jelölő üres soroknak viszont üres soroknak kell maradniuk, vagyis oda nem kerülhet semmi.

A nevek kategóriákba való sorolásánál a *NerKor annotálási útmutató*-ban lefektetett elveket követjük.

3. Morfológiai annotáció

A korpusz egy része morfológiai annotációt is fog tartalmazni. Kétféle morfológiai annotációt fogunk előállítani: megtartjuk az `emMorph` által kiadott elemzést, valamint átkonvertáljuk azt a UD jelenleg elérhető aktuális morfológiai formalizmusára (v2).

3.1. Az `emMorph` formalizmusa

Az `emMorph` teljes morfológiai elemzést ad ki, ami magában foglalja a lemmát, a szófajkódot és a morfoszintaktikai információkat is, például:
`adtad ad [/V][Pst.Def.2Sg]`

A lemma kerül a `LEMMA` oszlopba, a szögletes zárójelekben levő teljes címke pedig az `XPOS` oszlopba.

A címkék feloldása példákkal együtt az alábbi oldalon található:
`https://e-magyar.hu/hu/textmodules/emmorph_codelist`

3.2. A UD formalizmusa

A UD esetében a szófajkódoknak egy véges listája létezik[4] – ezek közül valamelyik kerül a `UPOS` oszlopba. A morfoszintaktikai információkat linearizált jegy–érték párok alkotják, melyek között vannak univerzálisak[5] és nyelvspecifikusak is[6], például:

```
VERB
Definite=Def|Mood=Ind|Number=Sing|Person=2|Tense=Past|VerbForm=Fin
|Voice=Act
```

Ezek a jegyek kerülnek a `FEATS` oszlopba.

Hivatkozások

[Novák et al., 2016] Novák, A., Siklósi, B., and Oravecz, Cs. (2016). A New Integrated Open-source Morphological Analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

[Tjong Kim Sang, 2002] Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In Roth, D. and van den Bosch, A., editors, *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

[Tjong Kim Sang and De Meulder, 2003] Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003*. Edmonton, Canada.

[4] `https://universaldependencies.org/u/pos/index.html`
[5] `https://universaldependencies.org/u/feat/index.html`
[6] `https://github.com/dlt-rilmta/panmorph`