Osasun-arloko entitate izendunen etiketatzea

1. Proposatzailea: Maite Oronoz
2. Deskribapena

Hizkuntzaren prozesamenduko eskoletan ikusi ditugun entitate izendunen etiketatzaileek (NER edo named entity recognizers), orokorrean pertsona-izenak, leku-izenak eta erakunde-izenak ezagutu ohi dituzte .

Osasun-txostenetan gaixotasunak, botika-izenak, sintomak etab. ezagutu ohi dira. Ataza honi Clinical named entity recognition egitea deritzo.
Adibidez, demagun testu hau dugula:

```
DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis
```

Bertan, elementu hauek ezagutuko lirateke:

```
[DCTN4] as a modifier of [chronic Pseudomonas aeruginosa infection] in [cystic fibrosis]
```

Lan honetan MedMentions corpusa erabili nahi dugu (train, dev eta test) zatiak ditu:

https://github.com/chanzuckerberg/MedMentions

Corpus hau eskuz etiketatu dute UMLS (Unified Medical Language System) etiketekin. Hau da goiko adibidean, osasun-alorreko terminoak identifikatzeaz gain, zein kontzeptu identifikadoreri lotzen zaizkion (CUI, Concept Unique identifier) adierazten da (azken zutabea da CUI identifikadorea):

0	5	DCTN4	C4308010
23	63	chronic Pseudomonas aeruginosa infection	C0854135
67	82	cystic fibrosis	C0010674

Lan honetan oinarrizko hurbilpena eta hurbilpen egokitua landuko dira.

3. Helburuak

Helburu orokorra medikuntzako NER sistema gainbegiratu bat egitea da. Sistemak train zatian ikasiko du, dev-en probatuko duzu, eta azkenik, test-eko emaitzak erakutsiko dituzu.

Horretarako, kasu guztietan lehenengo hau egin behar da: **MedMentions-eko formatua BIO etiketatzera (edo beharrezkoa den beste formaturen batera) egokitu (train, dev eta test)** (kontuz! lehenengo aukeratu erabiliko duzun sistema, sistema guztiek ez baitute eskatzen sarreran BIO formatua). **Ezinbestekoa da test multzoaren kontra ebaluazioa**

egitea eta eskolan ikusitako metrikak taula batean ematea (estaldura, doitasuna eta f-measure gutxienez).

Helburuak zailtasun mailaren araberakoak izango dira:

- Z1 (minimoa): MedMentions corpusa erabiliz NER sistema orokor bat entrenatu, termino bat UMLSkoa den, hau da, osasunarekin zerikusia ote duen jakiteko. Hau da, terminoak identifikatu, klase bakarra irteeran duzularik (Medikuntzkoa edo MED).

  `[DCTN4] as a modifier of [chronic Pseudomonas aeruginosa infection] in [cystic fibrosis]`

  Eskolan ikusitako NER sistema bat erabil dezakezu.

- Z2 (ertaina): Hizkuntza-eredu bat doitu (*fine-tuning*) MedMentions-eko corpusarekin NER egiteko.
  Sistemari hobekuntzak egitea hobesten da. Aurrekoan bezala, sistema test multzoaren kontra ebaluatu egin behar da eta ebaluazio-metrikak eman beharko dira taula batean.
  Klase bakararekin lan egiteaz gain, Medmentions-eko klase semantiko ezberdinak bereiztea ongi legoke (T klaseak, adibidez, “Virus”, “Cell”, “Antibiotic”….)
  Klaseak hemen.

- Z3 (altua):

Eskolan ikusi ez dugun NER sistema bat erabili. Gainera, Z2 atala hobetu daiteke esperimentuekin (klase kopuruarekin jokatzen, edo granularitatearekin jokatzen, adibidez) edo/eta teknika ezberdinak erabil daitezke ebaluazio-metrikak hobetzeko… Adibidez:

- Osasun-arlorako NER sistema espezifiko bat erabili (batzuk hauei CliNER deitzen diete).
- edo MetaMap eta zuk entrenatutako NERen bat erabili (Z1 edo Z2koa izan daiteke) sistemak zuk nahi duzun moduan konbinatuaz.

Ebaluazioa modu berean.
Sistema hobetzeko moduak ikaslearen esku geratzen dira.

4. Materialak

Proiektu honetarako materialak hurrengoak dira:

- Z1, Z2 eta Z3: MedMentions Corpusa: https://github.com/chanzuckerberg/MedMentions
  Eskolan ikusitako NER sistema bat
- Z2: Hizkuntza-eredu bat doitu NER egiteko.
- Z3: Hauetakoren bat?
  - https://metamap.nlm.nih.gov/
  - https://www.nlm.nih.gov/research/umls/index.html
  - Medical edo Clinical NER sistema bat

MetaMap edo UMLS erabiltzeko, “National Library of Medicine”-n kontua behar da. Kontu pertsonala egin dezakezu (nahiko zaila urtero inkesta bat bete behar baituzu) edo “Research Organization” aukera erabil dezakezu eta bertan “University of the Basque Country” aukeratu.

5. Erreferentziak

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

Sunil Mohan and Donghui Li. *MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts*. In Proceedings of the 2019 Conference on Automated Knowledge Base Construction (AKBC 2019). Amherst, Massachusetts, USA. May 2019. Preprint: https://arxiv.org/abs/1902.09476