Named Entity Annotation II

21-28 Temmuz 2017

Problem: Cümlede named-entity (özel isim) bulma.

Araç: http://brat.cognitus.ai/#/
Başlangıçta incelemek iyi olur: http://brat.nlplab.org/manual.html

Diğer guidelinelar:

- Conll03 Shared Task: http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt
- MUC 1997: http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html
- Europarl: http://web.mit.edu/andreeab/www/corpus/annotationGuidelines.pdf

+ Bu kaynaklardaki ve başka bazı sistemlerden kuralları da kendi kural setimize ekledik; fakat hepsini değil, bazılarını kendi ihtiyaçlarımıza göre değiştirdik.

Etiketler

PER

➢ Person. Insan isimleri. Eğer birden fazla kelime varsa, Ernesto Che Guevara gibi, yalnız Ernesto'yu değil veya tek tek isimleri değil, bütün olarak Ernesto Che Guevara'yı etiketleyelim.

Bütün insan isimleri

- Kısaltmalar dahil.
- Yerleşik, ekstra entity içermeyen ünvan-hitapları dahil edebilirsiniz
    - PER<Ayşe hanım>, PER<Prof. Dr. Ayşe>, PER<Başbakan Hasan Hüseyin>, PER<Doktor Hikmet>
        - Devamında yalnızca ünvan geçiyorsa (Doktor hastaneye geldi), isimsiz ünvanları almayalım.
        - Futbolcu, sınıf başkanı vs gibi yaygın olmayan, onlarsız da ismin geçebildiği ünvanları etiketlemeye katmayalım.
    - Ünvanda kurum ismi geçiyorsa, ünvan MISC, insan ismi de PER
        - Demokratik Suudi Arabistan Cumhuriyeti Eşbaşbakanı Fatma bizi ziyaret etti -> MISC<Demokratik Suudi Arabistan Cumhuriyeti Eşbaşbakanı> PER<Fatma> bizi ziyaret etti.

Kurgusal isimler
- PER<Bay K.>

Hayvan isimleri

Takma isimler

LOC

➢ Location. Yer isimleri. Organizasyondansa mekan olarak bilinen / tanınan her isim buna dahil, X Hipodromu, X stadyumu, X kampüsü.. Bütün coğrafi kullanımlar.

Yerleşim birimleri (kıta, ülke, bölge, şehir, ilçe, kasaba, köy)
- İstanbul'un Kadıköy kasabasında xxx.. -> LOC<İstanbul'un> LOC<Kadıköy> kasabasında xxx.. Diye ayrı ayrı işaretleyelim fakat "kasabasında" kelimesini dahil etmeye gerek yok çünkü semt/kasaba ve daha büyük yerleşim yeri isimleri genellikle tek başlarına anılır (Yeldeğirmeni semtinde yerine Yeldeğirmeni'nde..). Fakat:
    - İstanbul'un Kadıköy ilçesinin Yeldeğirmeni semtinde Rasimpaşa Mahallesi'nde xxx -> LOC<İstanbul'un> LOC<Kadıköy> ilçesinin LOC<Yeldeğirmeni> semtinde LOC<Rasimpaşa Mahallesi'nde> xxx. Mahalle ve daha alt adres ifade eden isimleri etiketlemeye dahil edelim.

Adresler (mahalleler, caddeler, yollar, sokaklar)
- Ayrı ayrı işaretleyelim. Davutpaşa Mh. Zeytinburnu Cd. -> LOC<Davutpaşa Mh.> LOC<Zeytinburnu Cd.>
- Kapı vs numaralarını işaretlemeye gerek yok.

Doğal yerler
Dağlar, nehirler, denizler, göller, vadiler, zirveler, gök cisimleri
- Isimle birlikte (LOC<Everest Tepesi> veya LOC<Everest tepesi>)

Doğal olmayan açık/kapalı yerler Meydanlar, pazarlar, avm'ler, okullar, tiyatrolar, sinemalar, hastaneler, parklar, spor salonları, köprüler, limanlar, barajlar, plajlar, kütüphaneler, kampüsler
İsimli binalar Oteller, hosteller, yurtlar, kiliseler, camiler, dini yerler, galeriler, merkezler, kafeler, lokantalar
Soyut, kurgusal yerler - LOC<Kaf Dağı>

ORG

➢ Organization. Tüm kurum, örgüt, yapı isimleri.

Sivil yapılar Partiler, örgütler, sendikalar - İç içe yapı isimlerine dikkat edelim: - DİSK Dev Yapı İş Sendikası -> ORG<DİSK> ORG<Dev-Yapı İş Sendikası> Topluluklar, spor kulüpleri, takımlar Müzik, sanat grupları
Idari yapılar Bakanlıklar, hükümetler (ORG<XX hükümeti>), müdürlükler, temsilcilikler (ORG<ABD Konsolosluğu>), yargı birimleri (ORG<İstanbul Cumhuriyet Başsavcılığı>), birlikler (ORG<AB>, ORG<BM>)
Yayınlar Gazeteler, dergiler - ORG<XX Gazetesi>, ORG<XX dergisi>
Ticari Şirketler, bankalar, fabrikalar

MISC

➢ Miscellaneous. Ne PER, ne ORG, ne LOC diyebildiğiniz ama özel isim olduğunu bildiğiniz herşey.

Diller, dinler, milliyetler, ideolojiler, akımlar
- MISC<Afrikalı>, MISC<Çinli>, MISC<Arap>
- MISC<Bayburtlu>, MISC<AKP'li>
- MISC<Marksizm>
  - Ideolojinin taraftarları, topluluk halindeyse genellikle MISC; bağlama göre karar verebilirsiniz.
  - MISC<Milliyetçiler> seçimi kaybetti.
  - Takım taraftarları
  - MISC<Beşiktaşlılar>

Çağlar, dönemler - Tunç Devri, Neolitik Çağ, 60'lar..
Savaşlar, çatışmalar
Sloganlar Başlıklar - Film, kitap, oyun, şarkı   - MISC<TC anayasası>
Etkinlikler, tarihsel olaylar - XX Festivali, XX yarışı, XX konseri..   - LOC<İstanbul Film Festivali> : istanbul'u ayrıca LOC olarak ayırmayalım. - MISC<31 Mart Vakası>, MISC<Paris Komünü>, MISC<İran devrimi>
Belirli günler ve haftalar - MISC<Dünya Kadınlar Günü> veya MISC<8 Mart Dünya Kadınlar Günü> : tarihi dahil edelim. - MISC<Kanlı Pazar>
Seriler, tipler, ürünler, ilaç isimleri - MISC<F16> - MISC<Samsung C9>
Canlıların biyolojik isimleri - MISC<Lychnis coronaria> - MISC<Felis catus>

Önemli noktalar

+ LOC | ORG
- Mekandan bağımsız, yapıya atfen kullanım varsa ORG olarak karar verelim. Coğrafi ise LOC diyelim.

    + ABD’nin kararını tanımıyoruz. -> ORG<ABD>
    + ABD’ye gitmiş. -> LOC<ABD’ye>

    - Etkinliği XX Tiyatrosu düzenliyor -> ORG<XX Tiyatrosu>
    - Oyun XX Tiyatrosu’nda oynuyor -> LOC<XX Tiyatrosu’nda>

+ LOC | MISC
    - Mekanın ismi başka birşeye, film, kitap ismine dönüşmüşse MISC seçelim.
        + LOC<Casablanca’yı> izledik.

+ Canlı isimleri
    - Canlılara insanların bilimsel bağlam dışında verdikleri isimler -> PER
    - Bilimsel isimler -> MISC

+ Farklı dillerden entity’lerle karşılaşırsanız etiketleyebilirsiniz.
    + MISC<Moonlight Sonata>

+ Yanlış yazımları veya türkçeleştirilmiş kelimeleri doğrusunu kabul edip etiketlemeye devam edebilirsiniz - bir düzeltme yapmanıza gerek yok.
    - Etya’da çalışıyorum -> ORG<Etya>
    - Vaşington’a gitmiş. -> LOC<Vaşington>

+ Yapı içinde yapı belirtildiğinde ayrı ayrı etiketleyelim:
    - Orta Doğu Teknik Üniversitesi Beşeri Bilimler Fakültesi Sosyoloji Bölümü’nden konuşmacılar.. -> ORG<Orta Doğu Teknik Üniversitesi> ORG<Beşeri Bilimler Fakültesi> ORG<Sosyoloji Bölümü’nden> konuşmacılar..
        - Tabii ki Orta Doğu’yu ayrıca LOC diye işaretlemiyoruz :)

+ Entity’lerin başındaki sıfatları seçmenize gerek yok, yalnızca entity’i alalım.
    - Eski Sovyetler Birliği -> Eski LOC<Sovyetler Birliği>

+ PER’in ünvanını tanımlayan kurum isimlerini ayrı etiketleyelim
    - MISC ve ORG farkına dikkat.
        + Marmara Üniversitesi Rektörü Berke Can -> MISC<Marmara Üniversitesi Rektörü> PER<Berke Can>
        + Marmara Üniversitesi Rektörlüğüne kayyum atandı. -> ORG<Marmara Üniversitesi Rektörlüğüne>

➢ Her büyük harfli kelime özel isim olmayabilir ;)
➢ Emin olmadığımız noktaları not edip üstünden geçelim, yazışalım, birbirimize soralım.

- ➢ Kelimelerin bağlamlarına dikkat edelim.
- ➢ Mesela Beşiktaş hem LOC, hem ORG olabilir. Cümledeki kullanıma göre karar vermeliyiz.
- ➢ Çakışmalardan kaçınalım. Yani hem yıldız teknik üniversitesi'ni ORG diye, hem de yıldız'ı LOC diye işaretlemeyelim. Cümlede önemli olan yıldız semti değil, yıldız teknik üniversitesi'dir.
- ➢ Kesme işaretinden sonraki eki seçebilirsiniz. (Yıldız'da kelimesi için tamamını seçmek sorun değil.)
- ➢ Rekürsif etiketlemeden kaçınalım:
    - ○ Marmara Üniversitesi Rektörü Berke Can gibi bir ifadede, hem ORG<Marmara Üniversitesi> hem MISC<Marmara Üniversitesi Rektörü> gibi bir etiketleme yapmayalım; burada doğru olan yalnızca MISC.