**Bu veriyi bir yayın çalışmasında kullacaksanız lütfen aşağıdaki yayına atıf yapınız:**

* U Orhan, E Arslan. Learning Word-Vector Quantization: A Case Study in Morphological Disambiguation, Transactions on Asian and Low-Resource Language Information Processing, 19, 5, 72, 2020.

**VERİSETİ HAKKINDA**

Bu veriseti, 26.03.2019 ve 13.04.2019 tarih aralığındaki internet haber yayınlarından seçilmiş metinler ve 3 adet epub formatındaki roman içerik kullanılarak oluşturulmuştur. Tüm metinlerden noktalama işaretleri temizlenmiş ve düzensiz boşluklar düzenlenmiştir. Bu işlemi takiben her metin cümlelere ayrılmış ve bu cümleler ile birlikte tüm metinler bir Mongo veritabanına aktarılmıştır. Bir Morfolojik analizcisi kullanılarak (Yıldız vd., 2019\*) cümleler birimlerine ayrılmış ve parse edilerek her kelime için analizcinin çıktıları Mongo veritabanında JSON formatında ilgili alana eklenmiştir. Etiketleme yapan dil uzmanı, hazırlanmış olan web tabanlı bir etiketleme programı yardımıyla bu Mongo veritabanına erişmiş ve Türkçe için kullanılan POS etiketlerinden (daraltılmış) seçim yapmıştır. Uzman, bazı bağımsız kelimeleri ekranda birleştirebilmekte (mwe) ve bu birleştirilmiş kelimelere gerekli etiketleri atayabilmektedir. Ayrıca ekranda karşısına gelen maddebaşı önerilerindeki yazım hatalarını elle düzeltmekte ve cümle içindeki durumuna göre belirsizlik içeren kelimelerin doğru olan POS etiketini seçebilmektedir.

**VERİSETİ İSTATİSTİKLERİ**

|  |  |  |
| --- | --- | --- |
|  | **Toplam** | **Açıklama** |
| **Cümle sayısı** | 5.723 |  |
| **Metin sayısı** | 219 | Haber: 216. Romanların her biri tek metin olarak işlenmiştir. |

**JSON YAPISI**

|  |  |  |
| --- | --- | --- |
| **JSON Ana Özelliği** | **JSON Alt Özelliği** | **Açıklama** |
| Cumle | N/A | Etiketlenecek olan cümle |
| IsTagged | N/A | Cümlenin etiketlenme durumu. ‘true’ etiketlenmiş, ‘false’ etiketlenmemiş |
| Text\_ID | N/A | Etiketlenecek olan cümleyi içeren metnin ID’si |
| Tokens | token | Etiketlenecek olan cümleden seçilen bir ‘birim’ |
|  | lemma | Birim’in maddebaşı |
|  | tag | Birim için etiketleyici kişi tarafından seçilen POS etiketi |
|  | flag | ‘Morfolojik analizcisi’ çıktısı.  0 -> Analizci tek çıktı sağlıyor. Belirsizlik yok  1 -> Analizci birden fazla çıktı sağlıyor. Belirsizlik var. Lemma adaylar ‘#’ ayıracı ile ayrılmıştır.  2 -> Analizci sözlük dışı (OOV) gibi bir durumdan kaynaklı olarak hiçbir çıktı sağlamamakta |

* Bu çalışma için önemsiz olan rakamların Postag’i etiketleyici tarafından PUNC olarak seçilmiştir.
* Oluşturulan bileşik kelimeler için flag=3 olarak seçilmiş ve ‘Noun’ ve ‘Verb’ olacak şekilde iki farklı POS ile etiketlenmiştir.
* Etiketlenmemiş cümlelerde boş kalmaması açsından tag alanı ‘DB’ olarak işaretlenmiştir.

\* OT Yıldız, B Avar, G Ercan. An open, extendible, and fast Turkish morphological analyzer, 2019

**BAZI ÖRNEKLER**

Aşağıdaki örnekte etiketleyici ‘bir’ birimi için ‘ADJ’ yani sıfat seçimi yapmıştır. Ancak Morfolojik analizcinin sunduğu seçeneklerde sadece ‘DET’ ve ‘NOUN’ POS etiketi bulunmaktadır.



Aşağıdaki örnekte etiketleyici ‘faltaşı birimi için ‘NOUN’ yani isim seçimi yapmıştır. Ancak Morfolojik analizci bu birim için seçenek sunmamıştır.

