Yapay Sinir Ağları ile Türkçede Konuşma Etiketleme (Pos Tagging) Yapabilmek için Yöntem Geliştirme Örnek Tasarımı

Hazırlayan

Fırat Kaan Bitmez

Öğrenci Numarası

23281855

Dersin Hocası

Asst.Prof.Dr. İsmail İşeri

Giriş

Bu Raporda/Makalede Türkçe metinler üzerinde **Part-of-Speech (POS) tagging** yapmak için yapay sinir ağı kullanımını ele almaktadır. POS tagging her kelimenin dil bilgisi açısından rolünü belirlemeyi amaçlayan bir işlemdir. Bu çalışmada, Türkçe metinlerden oluşan bir veri kümesi kullanılarak bir yapay sinir ağı örnek model tasarımı ve aşamaları oluşturulmuştur.

Amacımız

Bu çalışmanın amacı, Türkçe metinlerde **Yapay sinir ağı** kullanılarak Türkçe Pos Tagging yapabilmek için veri kümesinin nasıl hazırlanmasından başlayarak, özellik vektörünün yapısının nasıl olabileceği konusunda bir örnek tasarım hazırlamak ve tasarım örneğini hazırlarken daha önce yapılan çalışmalardan olan **DEVELOPING METHODS FOR PART OF SPEECH TAGGING IN TURKISH** isimli makaleyi inceleyeceğiz nasıl adımlar izlediklerini çıkarım yapacağız daha sonra ise biz Yapay Sinir Ağı kullanarak bu adımları nasıl yaparız sorusuna cevaplar arayacağız. İnceleyeceğimiz makalenin linkini en altta **kaynaklar** bölümünde bulabilirsiniz.

DEVELOPING METHODS FOR PART OF SPEECH TAGGING IN TURKISH Makalesinden Kısaca Bahsetmemiz Gerekirse:

Bu makalede Türkçe konuşma Etiketlemede **SVM**(Destek Vektör Makineleri) Yaklaşımı kullanılarak örnek bir yöntem geliştirilmiş. Tabi daha öncesinde Part-of-Speech, Pos Tagging gibi. kavramlara geniş bir yer verilerek bu kavramlar açıklanmış. Kısaca bahsetmem gerekirse:

Part-of-Speech Tagging = Konuşma sırasında Etiketleme

Part-of-Speech Tagging: Bir metindeki sözcüklerin belirli bir şeye karşılık gelecek şekilde etiketlenmesi POS etiketleme morfolojik analizin basitleştirilmiş bir biçimi olarak kabul edilebilir.

Neden part-of-speech Tagging'e ihtiyacımız var? : Konuşmanın bir kısımını etiketleme ile adımlardan biridir. Doğal Dil İşlemede önceki aşamalar bir sonraki aşamayı besler bu yüzden ihtiyacımız var.

Part-of-Speech Tagging Yaklaşımları

1) Kural Tabanlı Yaklaşım
2) TBL (Dönüşüm Temelli Öğrenme)
3) Markov Modeli Yaklaşımları
4) Maximum Entropi Yaklaşımları
5) Destek Vektör Makineleri (SVM)

DEVELOPING METHODS FOR PART OF SPEECH TAGGING IN TURKISH
Makalesinde yapılan çalışmada kullanılan **SVM** yaklaşımıdır.

SVM, diğer yöntemlere göre iki temel avantaja sahiptir. Yüksek boyutlu uzaylarda işlem yapabilme yeteneğine sahiptir ve bu, genellikle birçok özelliğin kullanılması durumunda yararlıdır. Ayrıca, aşırı uymaya karşı dayanıklıdır, yani genelleme yeteneği daha iyidir.

Destek Vektör Makineleri (SVM) ve Yapay sinir ağları (YSA) makine öğrenmesinin farklı alanlarında kullanılan iki önemli algoritmadır. İşlevselliklerinde ve uygulama alanlarında bazı benzerlikler bulunmasına rağmen, birçok açıdan farklılık gösterirler.

Biz SVM yerine YSA algoritmasını kullanarak Türkçede pos tagging işleminin yapılması için nasıl bir model oluşturmamız gerektiğini anlamaya çalışacağız ve örnek bir tasarım yapacağız.

Örnek Model Tasarım Aşamaları

Veri Kümesi Hazırlığı:

1) **Veri Toplama**: İlk adım veri toplama, POS etiketli Türkçe cümleler içeren bir veri kümesi oluşturmaktır. Bu veri kümesi, farklı türlerde metinlerden (makaleler, köşe yazıları, kitaplar vb.) toplanabilir.

2) **Veri Temizleme ve Ön İşleme**: Toplanan veri kümesi, gereksiz karakterlerin ve boşlukların temizlenmesi için işlenmelidir. Burada aslında temizle işlemine gürültü, kirlilik gibi farklı isimlerde verilebiliyor. Ardından, cümleler ayrıştırılmalı ve kelimelerine ayrılmalıdır.

3) **POS Etiketleme**: Her kelimeye uygun bir POS etiketi atanmalıdır. Bu etiketler, veri kümesinde bulunan POS etiketleriyle uyumlu olmalıdır.

Özellik Vektörünün Yapısı:

1) **Kelime Özellikleri**: Her kelimenin kendisine özgü özellikleri vardır.
   Örneğin:
   - Kelimenin uzunluğu
   - Baş harfinin büyük/küçük olması
   - Noktalama işareti içermesi
   - Rakam içermesi

2) **Önceki ve Sonraki Kelimelerin Özellikleri**: Bir kelimenin POS etiketi, önceki ve sonraki kelimelerin POS etiketlerine bağlı olabilir.
   Örneğin:
   - Önceki kelimenin POS etiketi
   - Sonraki kelimenin POS etiketi

- Önceki kelimenin son üç karakteri
- Sonraki kelimenin ilk iki karakteri

3) **Sözlük Bazlı Özellikler**: Türkçe dil bilgisi kurallarını temsil eden bir sözlük oluşturulabilir. Bu sözlük, belirli kelimelerin belirli POS etiketleriyle ilişkilendirilmesini içerir.
   Örneğin:
   - Kelimenin Türkçe sözlükte bulunup bulunmaması
   - Kelimenin sözlükteki anlam sayısı
   - Kelimenin yaygınlığı

Yapay Sinir Ağı Tasarımı

1. Giriş Katmanı:

Giriş katmanı, metindeki her kelimenin özelliklerini temsil eden bir vektör alır.

Her kelimenin özelliklerinin bir araya getirilmesiyle oluşturulan vektör, yapay sinir ağının girişine verilir.

2. Gizli Katmanlar:

Yapay sinir ağında bir veya daha fazla gizli katman bulunabilir.

Her gizli katmanda, giriş vektöründeki özelliklerin bir ağırlık matrisiyle çarpılması ve bir aktivasyon fonksiyonu ile işlenmesi gerçekleştirilir.

Örneğin, **ReLU** (Rectified Linear Activation) veya **sigmoid** aktivasyon fonksiyonları sıklıkla kullanılır.

3. Çıkış Katmanı:

Çıkış katmanı, POS etiketlerinin olasılıklarını içerir.

Birden fazla POS etiketi varsa, çoklu sınıflandırma için **softmax** aktivasyon fonksiyonu kullanılabilir.

Çıkış katmanında, her bir POS etiketinin olasılığını belirleyen bir düğüm bulunur.

4. Eğitim:

Yapay sinir ağı, eğitim veri kümesi üzerinde eğitilir.

Gerçek çıktılarla (POS etiketleri) tahmin edilen çıktılar arasındaki hatayı minimize etmek için bir kayıp fonksiyonu kullanılır.

Kayıp fonksiyonu olarak, çapraz-entropi kaybı sıklıkla tercih edilir.

5. **Değerlendirme:**

   Eğitim sona erdikten sonra, yapay sinir ağı test veri kümesi üzerinde değerlendirilir.

   Doğruluk ve diğer metrikler kullanılarak modelin performansı değerlendirilir.

Bu tasarım, yapay sinir ağı modelinin yapısal bileşenlerini ve işleyişini açıklar. Giriş katmanı, gizli katmanlar ve çıkış katmanı, modelin veriyi işleme ve çıktı üretme şeklini tanımlar. Eğitim süreci, modelin belirli bir görevi öğrenmesini ve optimize etmesini sağlar. Son olarak, modelin performansı değerlendirilerek, ne kadar etkili olduğu belirlenir. Bu tasarım, Türkçe metinler üzerinde POS etiketleme görevi için bir çerçeve sağlar ve genişletilebilir veya iyileştirilebilir.

Kütüphaneler ve Araçlar

Python programlama dili kullanılarak **yapay sinir ağı örnek tasarım modeli** geliştirilmiştir.

Yapay sinir ağı modeli oluşturmak için **TensorFlow** veya **PyTorch** gibi **derin öğrenme kütüphaneleri** kullanılabilir.

Veri işleme ve özellik mühendisliği için **pandas**, **NumPy** ve **scikit-learn** gibi **kütüphaneler** kullanılabilir.

Metin verilerini işlemek için **NLTK** (Natural Language Toolkit) gibi doğal dil işleme kütüphaneleri kullanılabilir.

Örnek bir Python kod parçası yazarsak şu şekilde olabilir:

```
import numpy as np
import pandas as pd
import tensorflow as tf
from tensorflow.keras import layers, models
from sklearn.model_selection import train_test_split
from nltk.tokenize import word_tokenize
from nltk import pos_tag

# Bu alanda Veri kümesini yükleme ve ön işleme
# Veri kümesinin yüklenmesi, temizlenmesi ve POS etiketlerinin alınması işlemleri burada gerçekleştirilir.
# Özellik vektörlerinin oluşturulması
# Her kelime için özellik vektörleri oluşturulur. Bu özellik vektörleri, kelimenin uzunluğu, büyük/küçük harf olması gibi özellikleri içerir. Ama bizim amacımız bu tasarımda YSA aşamasına odaklanmak olduğu için bu adımları atlayacağız.
```

```
# Yapay sinir ağı (YSA) modelinin tanımlanması
model = models.Sequential([
    layers.Dense(128, activation='relu', input_shape=(feature_vector_size,)),
    layers.Dense(64, activation='relu'),
    layers.Dense(num_classes, activation='softmax')
])

# Modelin derlenmesi
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

# Modelin eğitilmesi
history = model.fit(X_train, y_train, epochs=10, batch_size=32, validation_data=(X_val, y_val))

# Modelin değerlendirilmesi
test_loss, test_accuracy = model.evaluate(X_test, y_test)
```

Sonuç

Sonuç oalarak bu çalışmada, Türkçe metinler üzerinde doğru POS etiketlerini tahmin etmek için bir yapay sinir ağı model örnek bir tasarımı geliştirilmiştir. Veri hazırlığı, özellik vektörünün tasarımı ve modelin eğitim süreci gibi aşamalar hakkında bilgi sahibi olunmuştur. Geliştirilen modelin, test veri kümesi üzerindeki performansı test adımları yapılmamıştır sadece örnek bir tasarımda neler yapılmalı bunlar ele alınmıştır. Yapay sinir ağı modeli, Türkçe metinlerde POS etiketleme görevini etkili bir şekilde gerçekleştirebilecek güçlü bir araç/algoritma olarak öne çıkmaktadır. Bizde bu algortimayı nasıl kullanırız sorusuna cevap aramış olduk.

Bu çalışma, Türkçe doğal dil işleme alanında önemli bir adım olup, gelecekteki çalışmalar için bir temel oluşturmuştur. Modelin daha da iyileştirilmesi ve genişletilmesi, Türkçe metinler üzerindeki doğal dil işleme uygulamalarının kalitesini artıracaktır.

Öneriler

1) Modelin Performansını İyileştirmek İçin Daha Fazla Özellik Ekleyin: Özellik vektörü tasarımınızı genişletebilirsiniz.

2) Daha Fazla Veri Kullanın: Veri setinizi genişletmek, modelin genelleme yeteneğini artırabilir. Farklı kaynaklardan daha fazla POS etiketli Türkçe metinleri toplayabilir ve veri kümenizi çeşitlendirebilirsiniz.

3) Modelin Mimarisini Geliştirin: Modelinizin mimarisini ayarlayarak veya daha karmaşık bir yapıya geçerek performansı artırabilirsiniz. Örneğin, daha fazla gizli katman ekleyerek veya mevcut katmanların genişletilmesiyle modeli daha karmaşık hale getirebilirsiniz.

4) Transfer Learning Kullanın: Eğer elinizde yeterince büyük bir veri kümesi yoksa, transfer learning yöntemlerini kullanarak önceden eğitilmiş bir modelden başlayabilirsiniz. Örneğin, Türkçe metinler üzerinde genel doğal dil işleme görevlerinde önceden eğitilmiş bir dil modeli üzerine özelleştirme yapabilirsiniz.

5) Farklı Optimizasyon Algoritmalarını Deneyin: Modelinizi eğitmek için farklı optimizasyon algoritmalarını deneyerek daha iyi sonuçlar elde edebilirsiniz.

Kaynaklar

DEVELOPING METHODS FOR PART OF SPEECH TAGGING IN TURKISH

by Berna Arslan & Özlem Patan

https://www.cmpe.boun.edu.tr/~gungort/undergraduateprojects/Developing%20Methods%20for%20Part%20of%20Speech%20Tagging%20in%20Turkish.pdf