Veri Seti Hazırlama ve Teknik Analiz Raporu

Proje Başlığı: Entity Normalization & Context-Aware Spelling Correction

Veri Kaynağı: Wikipedia (English Corpus)

Hedef Hacim: 3,000 Nitelikli Cümle (İleride artabilir)

Rapor Tarihi: 18.02.2026

1. Amacı ve Kapsamı

Bu çalışma; doğal dil işleme (NLP) modellerinin, metin içerisindeki Özel Varlıkları (Entities) doğru tanımasını ve bu varlıkların yazım hatalarını bağlamsal (context-aware) olarak düzeltebilmesini sağlamak için yüksek kaliteli bir eğitim seti oluşturmayı hedefler. Wikipedia'nın seçilme nedeni, varlıkların (kişi, kurum, yer) dilbilgisi kurallarına en uygun ve zengin bağlamda kullanıldığı kaynak olmasıdır.

2. Veri Çeşitliliği ve Kategori Stratejisi

Modelin belirli bir alana sıkışıp kalmaması (overfitting) için veri seti 5 ana kategoride çeşitlendirilmiştir:

Kategori	Öne Çıkan Başlıklar	Amaç
Global Politika	NATO, Diplomacy, International Law	Resmi dil ve diplomasi terminolojisi.
Bilim & Teknoloji	AI, Physics, NASA, Internet	Teknik terimler ve karmaşık varlık isimleri.
Türkiye Odağı	Atatürk, Istanbul, Ottoman Empire	Yerel varlıkların global bağlamda temsili.
Liderler & Biyografi	Churchill, Einstein, Angela Merkel	Kişi isimleri ve unvanların (Title) korunması.
Kurumsal & Popüler	Microsoft, Coca-Cola, NBA	Marka isimleri ve modern kültürel varlıklar.

3. Veri İşleme Hattı (Pipeline)

Ham metin, sistem tarafından "Altın Standart" veri haline gelene kadar şu aşamalardan geçer:

A. Gürültüden Arındırma (Cleaning)

Regex motoru kullanılarak metindeki gürültüler (noise) temizlenir:

- Atıf Temizliği: [1], [citation needed] gibi akademik işaretçiler silinir.
- Meta Veri Ayıklama: Sayfa düzenleme linkleri ([edit]) ve teknik kodlar elenir.
- Tipografik Düzeltme: Çift boşluklar ve hatalı noktalama boşlukları normalize edilir.

B. Dilbilimsel Filtreleme (Validation)

Her cümle, modelin öğrenme kalitesini artırmak için şu filtrelerden geçmek zorundadır:

- Uzunluk Denetimi: 5 ≤ Kelime Sayısı ≤ 40. Kısa başlıklar anlamsızdır, çok uzun cümleler ise bağlamı dağıtır.
- Sentaktik Doğruluk: Cümle büyük harfle başlamalı ve uygun bir duraklama işaretiyle (., !, ?) bitmelidir.
- Varlık Yoğunluğu: İçerisinde URL veya teknik "Category:" ibaresi barındıran cümleler otomatik reddedilir.

4. Varlık Normalizasyonu (Entity Normalization) Katkısı

Bu veri seti, "Varlık Normalizasyonu" için kritik olan "Çevresel Kelime Örüntülerini" saklar.

Örnek Senaryo:> Ham Metin: "Albert Einstien was born in Germany." (Hatalı yazım: Einstien)

Bağlam: Cümle içerisinde "born", "Germany" ve "Physics" gibi anahtar kelimeler geçtiğinde, modelimiz Einstien kelimesini Albert Einstein varlığına (Q937 wikidata ID) %99 güvenle normalize edebilir.

5. Veri Yapısı ve Depolama (JSON Format)

Veri seti hem makine hem de insan tarafından okunabilir bir hiyerarşide saklanır. Her bir entry şu metadata bilgilerini içerir:

- text: Temizlenmiş ham cümle.
- source_page: Verinin çekildiği Wikipedia başlığı.
- word_count: Cümlenin kelime uzunluğu.
- collected_at: Zaman damgası (Veri tazeliği takibi için).

6. Teknik Metrikler ve Başarı Kriterleri

- Denge Faktörü: Her sayfadan en fazla 22 cümle alınarak "Dominant Konu" riski bertaraf edilmiştir.
- Tekrarsızlık: Python dictionary (set tabanlı) yapısı sayesinde mükemmel cümle benzerlikleri elenmiş, veri setinin %100 özgün olması sağlanmıştır.
- Hız ve Etik: time.sleep(0.25) ile Wikipedia sunucularına etik bir yaklaşım sergilenmiş, veri toplama süreci kesintisiz hale getirilmiştir.