

PROJE ANALİZ RAPORU

ENTITY NORMALIZATION & CONTEXT-AWARE SPELL CORRECTION

1. PROJE KONUSU

Bu proje, İngilizce haber akışındaki metinlerde yer alan özel isimler (Named Entities) ve genel İngilizce kelimelerdeki yazım hatalarının otomatik olarak tespit edilmesi ve bağlama duyarlı şekilde düzeltilmesini amaçlamaktadır.

Sistem, özellikle politik figürler, kurumlar, ülkeler ve coğrafi adlar gibi özel isimlerde meydana gelen deformasyonları dilin fonetik yapısına ve uluslararası standartlara uygun biçimde normalize ederken; aynı zamanda genel İngilizce kelimelerdeki tipografik ve ortografik hataları da bağlamı koruyarak düzeltmeyi hedeflemektedir.

2. PROBLEM TANIMI

Dijital haber üretim süreçlerinde hız baskısı, klavye uyumsuzluğu, çok dilli içerik üretimi ve otomatik veri akışları nedeniyle hem özel isimlerde hem de genel İngilizce kelimelerde ciddi yazım deformasyonları meydana gelmektedir.

Bu deformasyonlar:

- Haber içeriklerinin doğruluğunu ve güvenilirliğini azaltmakta,
- Arama motoru optimizasyonunu (SEO) olumsuz etkilemekte,
- Bilgi erişim sistemlerinde eşleşme hatalarına yol açmakta,
- Metin madenciliği ve veri analitiği süreçlerinde veri bütünlüğünü bozmaktadır.

Bunun yanı sıra, bu tür hatalar haber kaynağına duyulan güveni ve kaynağın sahip olduğu prestiji de sarsabilecek küçük ama medya sektöründeki firmalar için büyük etkiler doğuran itibar riskleri oluşturmaktadır.

Bu nedenle hem özel isim normalizasyonu hem de genel bağlama duyarlı yazım düzeltmesi yapabilen bütünlük bir sisteme ihtiyaç duyulmaktadır.

3. YAZIM HATALARI SINIFLANDIRMASI

Haber metinlerinde karşılaşılan yazım hataları dört temel kategoride ele alınmaktadır.

1) De-asciification (Karakter Dönüşümü)

Türkçe karakterlerin İngilizce klavye uyumluluğu nedeniyle ASCII karakterlere indirgenmesi veya yanlış dönüştürülmesi durumudur.

- Karakter Değişimi:** u→ü, o→ö, s→ş, g→ğ, c→ç gibi spesifik harf kayıpları.
- Noktalama ve Büyük/Küçük Harf Hataları:** İngilizce ve Türkçe arasındaki i→İ ve l→ı (noktalı/noktasız) karmaşası.
- Örnekler:** Türkiye → Türkiye, İstanbul → İstanbul, TURKIYE → TÜRKİYE.

2) Harf Dizilimi ve Tipografik Hatalar

- **Harf Eksikliği (Omission):** Karakterlerin metinden düşmesi (Örn: Erdogan → Erdgan, President → Prsident).
- **Harf Fazlalığı (Insertion):** Gereksiz karakter eklenmesi (Örn: Türkiye → Türkiyee, Political → Poliitcal).
- **Yer Değiştirme (Transposition):** Harflerin sırasının karışması (Örn: Süleyman → Sülyeman, Israeli → Irsaeli).
- **Yanlış Harf (Substitution):** Klavyede komşu olan veya benzer sesli yanlış harf kullanımı (Örn: Ankara → Abkara, Taliban → Takiban).
- **Birleşik/Ayrı Yazım Hataları:** Özel isimler arasındaki boşluk karakterinin kaybı (Örn: Abdullah Gül → AbdullahGül, New York → NewYork).

3) Uluslararası İsim Değişimleri (Terminoloji)

Ülkelerin veya kurumların uluslararası alanda talep ettiği resmi isim değişikliklerinin takibi ve eski kullanımların güncellenmesi sorunudur.

- **Örnekler:** Turkey → Türkiye (2022 sonrası), Burma → Myanmar, Swaziland → Eswatini.

4) Genel İngilizce Yazım Hataları (Lexical Spelling Errors)

Özel isim olmayan standart İngilizce kelimelerde görülen ortografik hatalardır.

- goverment → government
- enviroment → environment
- politcal → political
- recieve → receive

Bu hatalar bağlama duyarlı biçimde düzeltilmelidir. Çünkü bazı kelimeler doğru yazıldığında farklı anlamlara gelebilir veya isimlerle karışabilir.

4. PROBLEM KISITLARI (CONSTRAINTS)

Projenin başarısı için aşılması gereken temel teknik kısıtlar şunlardır:

1) Cümle Yapısının ve Bağlamın Korunması

İngilizce dil bilgisi yapısı bozulmamalıdır. Sistem yalnızca hatalı yazımları düzeltmeli, doğru kelimelere müdahale etmemelidir.

2) Seçici Müdahale (Entity vs Non-Entity Ayrımı)

Sistem, özel isimleri ve genel kelimeleri doğru şekilde ayırt ederek her biri için uygun düzeltme stratejisini uygulamalıdır.

3) OOV (Out-of-Vocabulary) Kapasitesi

Eđitim verisinde bulunmayan yeni kiři, kurum veya yer adları karakter yapısından ve fonetik benzerlikten yararlanılarak tanınabilmelidir.

4) Dile Bađlı Belirsizlik (Ambiguity)

Bazı kelimeler hem özel isim hem de genel kelime olabilir (May, Bill, Brown vb.). Sistem bađlamı analiz ederek dođru yorumu seřmelidir.

5) Over-Correction Riskinin Önlenmesi

Haber metinlerinde yanlış düzeltme ciddi anlam kaymalarına neden olabilir. Sistem gereksiz müdahaleden kaçınmalı ve yüksek güven eřiđi ile çalışmalıdır. Yanlış pozitif düzeltmeler, dođru yazılmış bir özel ismi bozma riski taşıdığından, sistemin hata toleransı düşük tutulmalıdır.

6) Çok Dilli Etkileşim

Metinler İngilizce olmakla birlikte Türkçe karakterler ve farklı dil kökenli isimler içerebilir. Sistem çok dilli karakter yapılarıyla uyumlu çalışmalıdır.

7) Tutarlılık ve Standartlaştırma

Aynı varlık metnin farklı bölümlerinde aynı biçimde normalize edilmelidir.

8) Gerçek Zamanlı İşlenebilirlik

Haber akışları yüksek hacimli olduğundan sistem ölçeklenebilir ve hızlı çalışmalıdır.

5. PROJENİN HEDEF ÇIKTISI

Geliştirilecek sistem:

- Metindeki hatalı özel isimleri dođru biçime normalize eder,
- Genel İngilizce yazım hatalarını bađlama duyarlı şekilde düzeltir,
- Cümle yapısını korur,
- Terminolojik standartları uygular,
- Veri bütünlüğünü ve metin güvenilirliğini artırır,
- Medya kuruluşlarının yayın kalitesi ve kurumsal itibarını korumasına katkı sağlar.

Sonuç olarak proje, haber metinlerinde yüksek dođruluklu otomatik metin standardizasyonu sağlayan bütünleşik bir dil işleme sistemi ortaya koymayı amaçlamaktadır.