

KUZGUNLAR

TÜRKÇE DOĞAL DİL İŞLEME



EKİP BİLGİSİ



Hüseyin ERDEM



Ankara Üniversitesi Bilgisayar Mühendisliği (2015-2020)
4 Yıl Robotik, 3 Yıl Yapay Zeka Deneyimi
Yapay Zeka Üzerine Coursera Sertifikaları

- [Neural Networks and Deep Learning](#)
- [Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization](#)
- [Structuring Machine Learning Projects](#)

Kaggle'da Doğal Dil İşleme üzerine çalışmalar

- [RoBERTa W/ Preprocessing on Tweet Sentiment Extraction dataset](#)

Hugging Face, BERT, roBERTa, RNN, LSTM, Attentions, Transformers, Pytorch konularında deneyim



Behçet ŞENTÜRK



Ankara Üniversitesi Bilgisayar Mühendisliği (2015-2020)
4 Yıl Robotik, 3 Yıl Yapay Zeka Deneyimi
Yapay Zeka Üzerine Edx Sertifikası

- [Data Science for Python](#)

Kaggle'da Doğal Dil İşleme üzerine çalışmalar

- [Roberta Q&A + NER in s-e word level](#)

Hugging Face, BERT, roBERTa, RNN, LSTM, Attentions, Transformers, Tensorflow konularında deneyim

PROBLEM NEDİR?

Veri Setleri

- Bağlamsal Anlama modellerinin (BERT, RoBERTa, Electra vb.) eğitilmesi için ham Türkçe veri eksikliği.
- NER görevleri için daha düzenli ve kapsamlı veri ihtiyacı.
- Mevcut Türkçe soru-cevap veri setinin konu kapsamının dar olması.

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	Base: 110 Large: 340	Base: 110 Large: 340	Base: 66	Base: ~110 Large: ~340
Training Time	Base: 8 x V100 x 12 days* Large: 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	Large: 1024 x V100 x 1 day; 4-5 times more than BERT.	Base: 8 x V100 x 3.5 days; 4 times less than BERT.	Large: 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	Base: 16 GB BERT data Large: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

PROBLEM NEDİR?

Veri Setleri

- Bağlamsal Anlama modellerinin (BERT, RoBERTa, Electra vb.) eğitilmesi için ham Türkçe veri eksikliği.
- NER görevleri için daha düzenli ve kapsamlı veri ihtiyacı.
- Mevcut Türkçe soru-cevap veri setinin konu kapsamının dar olması.

Automatically Annotated Turkish Corpus for Named Entity Recognition and Text Categorization using Large-Scale Gazetteers

H. Bahadir Sahin, Caglar Tirkaz, Eray Yildiz,
Mustafa Tolga Eren, Ozan Sonmez

Huawei Turkey Research and Development Center, Umraniye, Istanbul, Turkey
eray.yildiz@huawei.com

hbahadirsahin, caglartirkaz, tolgaeren, osonmez@gmail.com

Abstract

Turkish Wikipedia Named-Entity Recognition and Text Categorization (TWNERTC) dataset is a collection of automatically categorized and annotated sentences obtained from Wikipedia. We constructed large-scale gazetteers by using a graph crawler algorithm to extract relevant entity and domain information from a semantic knowledge base, Freebase. The constructed gazetteers contains approximately 300K entities with thousands of fine-grained entity types under 77 different domains. Since automated processes are

NER and TC are scarce. It is hard to manually construct datasets for these tasks due to excessive human effort, time and budget. In this paper, our motivation is to construct an automatically annotated dataset that would be very useful for NER and TC researches in Turkish.

The emergence of structured and linked semantic knowledge bases (KBs) provide an important opportunity to overcome these problems. Approaches that leverage such KBs can be found in literature (Heck et al., 2013; Gerber et al., 2013; Hoffart et al., 2011; Mendes et al., 2011). However, using the structured data from KBs is a challenging task for

PROBLEM NEDİR?

Veri Setleri

- Bağlamsal Anlama modellerinin (BERT, RoBERTa, Electra vb.) eğitilmesi için ham Türkçe veri eksikliği.
- NER görevleri için daha düzenli ve kapsamlı veri ihtiyacı.
- Mevcut Türkçe soru-cevap veri setinin konu kapsamının dar olması.

The screenshot shows a GitHub repository page for 'TQuad/turkish-nlp-qa-dataset'. The repository has 2 branches and 1 tag. The README.md file contains the following text:

```
README.md
Turkish NLP Q&A Dataset

| Türkçe Soru Cevap Veri Seti - Turkish Question Answering Dataset

Bu veri seti Teknofest 2018 Yapay Zeka yarışması kapsamında Türk & İslam Bilim Tarihi üzerine oluşturulan Türkçe Soru-Cevap veri setidir.
```

PROBLEM NEDİR?

Modeller

- Göreve özgü eğitilmiş Türkçe modellerin tümünün Bert tabanlı olması ancak Electra modelinin daha başarılı sonuçlar vermesi.
- NER görevleri için eğitilmiş modellerin çok az sayıda etiket sınıfı için tahmin yapabilmesi.

PoS tagging

The Turkish [IMST dataset](#) from Universal Dependencies is used for PoS tagging evaluation. We use the `dev` branch and commit `a6c955`. Result on development set is reported in brackets.

Model	Run 1	Run 2	Run 3	Run 4	Run 5	Avg.
ELECTRA small	(0.9567) / 0.9584	(0.9578) / 0.9589	(0.9564) / 0.9591	(0.9544) / 0.9585	(0.9545) / 0.9582	(0.9560) / 0.9586
ELECTRA base	(0.9707) / 0.9734	(0.9710) / 0.9734	(0.9712) / 0.9745	(0.9728) / 0.9719	(0.9711) / 0.9727	(0.9714) / 0.9732
mBERT	(0.9573) / 0.9580	(0.9554) / 0.9584	(0.9556) / 0.9591	(0.9594) / 0.9572	(0.9580) / 0.9586	(0.9571) / 0.9583
BERTurk (32k)	(0.9701) / 0.9712	(0.9731) / 0.9717	(0.9728) / 0.9730	(0.9719) / 0.9729	(0.9728) / 0.9708	(0.9722) / 0.9719
BERTurk (128k)	(0.9707) / 0.9732	(0.9716) / 0.9712	(0.9702) / 0.9722	(0.9675) / 0.9715	(0.9711) / 0.9729	(0.9703) / 0.9722
BERTurk uncased (32k)	(0.9707) / 0.9703	(0.9711) / 0.9713	(0.9715) / 0.9705	(0.9717) / 0.9719	(0.9718) / 0.9697	(0.9714) / 0.9707
BERTurk uncased (128k)	(0.9716) / 0.9726	(0.9715) / 0.9710	(0.9704) / 0.9720	(0.9715) / 0.9702	(0.9704) / 0.9693	(0.9711) / 0.9710
DistilBERTurk	(0.9648) / 0.9654	(0.9649) / 0.9642	(0.9654) / 0.9660	(0.9646) / 0.9650	(0.9637) / 0.9642	(0.9646) / 0.9650
XLM-RoBERTa	(0.9611) / 0.9620	(0.9629) / 0.9623	(0.9617) / 0.9602	(0.9602) / 0.9618	(0.9614) / 0.9629	(0.9614) / 0.9619

PROBLEM NEDİR?

Modeller

- Göreve özgü eğitilmiş Türkçe modellerin tümünün Bert tabanlı olması ancak Electra modelinin daha başarılı sonuçlar vermesi.
- NER görevleri için eğitilmiş modellerin çok az sayıda etiket sınıfı için tahmin yapabilmesi.

NER

NER dataset is similar to the one used in [this paper](#). We converted the dataset into CoNLL-like format and used a 80/10/10 training, development and test split. Result on development set is reported in brackets.

Model	Run 1	Run 2	Run 3	Run 4	Run 5	Avg.
ELECTRA small	(0.9447) / 0.9468	(0.9421) / 0.9439	(0.9421) / 0.9471	(0.9428) / 0.9434	(0.9439) / 0.9447	(0.9431) / 0.9452
ELECTRA base	(0.9564) / 0.9566	(0.9552) / 0.9557	(0.9579) / 0.9567	(0.9563) / 0.9570	(0.9568) / 0.9577	(0.9565) / 0.9567
mBERT	(0.9441) / 0.9420	(0.9448) / 0.9421	(0.9439) / 0.9421	(0.9444) / 0.9421	(0.9434) / 0.9436	(0.9441) / 0.9424
BERTurk (32k)	(0.9574) / 0.9550	(0.9534) / 0.9552	(0.9539) / 0.9570	(0.9550) / 0.9543	(0.9594) / 0.9531	(0.9558) / 0.9549
BERTurk (128k)	(0.9479) / 0.9494	(0.9569) / 0.9599	(0.9546) / 0.9571	(0.9549) / 0.9579	(0.9557) / 0.9534	(0.9540) / 0.9555
BERTurk uncased (32k)	(0.9529) / 0.9511	(0.9531) / 0.9520	(0.9533) / 0.9543	(0.9530) / 0.9522	(0.9523) / 0.9511	(0.9529) / 0.9521
BERTurk uncased (128k)	(0.9512) / 0.9531	(0.9502) / 0.9518	(0.9517) / 0.9520	(0.9513) / 0.9525	(0.9530) / 0.9546	(0.9515) / 0.9528
DistilBERTurk	(0.9418) / 0.9392	(0.9411) / 0.9415	(0.9382) / 0.9400	(0.9411) / 0.9427	(0.9417) / 0.9427	(0.9408) / 0.9412
XLM-RoBERTa	(0.9536) / 0.9541	(0.9517) / 0.9521	(0.9527) / 0.9530	(0.9493) / 0.9530	(0.9529) / 0.9516	(0.9520) / 0.9527

PROBLEM NEDİR?

Modeller

- Göreve özgü eğitilmiş Türkçe modellerin tümünün Bert tabanlı olması ancak Electra modelinin daha başarılı sonuçlar vermesi.
- NER görevleri için eğitilmiş modellerin çok az sayıda etiket sınıfı için tahmin yapabilmesi.

Published as a conference paper at ICLR 2020

Model	Train FLOPs	Params	SQuAD 1.1 dev		SQuAD 2.0 dev		SQuAD 2.0 test	
			EM	F1	EM	F1	EM	F1
BERT (ours)	7.1e20 (1x)	335M	88.0	93.7	84.7	87.5	—	—
ELECTRA-Base	6.4e19 (0.09x)	110M	84.5	90.8	80.5	83.3	—	—
ELECTRA-400K	7.1e20 (1x)	335M	88.7	94.2	86.9	89.6	—	—
ELECTRA-1.75M	3.1e21 (4.4x)	335M	89.7	94.9	88.0	90.6	88.7	91.4

Table 4: Results on the SQuAD for non-ensemble models.

PROBLEM NEDİR?

Modeller

- Göreve özgü eğitilmiş Türkçe modellerin tümünün Bert tabanlı olması ancak Electra modelinin daha başarılı sonuçlar vermesi.
- NER görevleri için eğitilmiş modellerin çok az sayıda etiket sınıfı için tahmin yapabilmeleri.

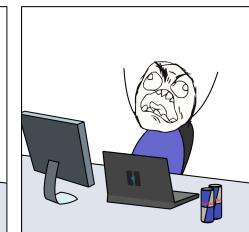
```
[1] 1 from transformers import pipeline, AutoModelForTokenClassification, AutoTokenizer  
2  
3 model = AutoModelForTokenClassification.from_pretrained("savasy/bert-base-turkish-ner-cased")  
4 tokenizer = AutoTokenizer.from_pretrained("savasy/bert-base-turkish-ner-cased")  
5 ner=pipeline('ner', model=model, tokenizer=tokenizer)  
6 ner.model.config.id2label
```

{0: 'B-LOC',
 1: 'B-ORG',
 2: 'B-PER',
 3: 'I-LOC',
 4: 'I-ORG',
 5: 'I-PER',
 6: 'O'}

PROBLEM NEDİR?

Kullanıcı Arayüzü

- Doğal dil işleme konusunda son kullanıcı veya hizmet sağlayıcı için kullanımı kolay bir ürün eksikliği
- Web sayfalarındaki uzun metinler içerisinde aranılan cevapları bulmanın zor olması.
- Eğitilmiş modellerin doğrudan son kullanıcı tarafından kullanılamaması.



ÇÖZÜM NEDİR?

Veri Setleri

- ~251 GB çevrimiçi pdf işlenilerek, Türkçe karakter hataları düzeltilmiş, 4 GB cümle veri seti hazırlanıldı.
- Var olan TWNERTC veri setinin 3076 farklı etiken içeren halinin 48 etiket sınıfına dönüştürülmesi ile daha düzenli bir NER veri seti elde edilmiştir
- Açık kaynak Türkçe soru-cevap veriseti olan TQUAD ile birlikte kullanılmak üzere 150 wikipedia içeriğinden 1135 soru cevap verisi hazırlandı.

ÇÖZÜM NEDİR?

Modeller

- Stefan Schweter tarafından ön eğitimi yapılmış (pretrained) ELECTRA base modeli üzerine NER, Soru-Cevap ve Duygu Durumu Analizi olmak üzere 3 farklı görev modeli fine-tune edilmiştir.
- Electra-base modeli, düzenlendiğimiz TWNERTC veri seti ile fine-tune edilmiştir. Bu model paylaşılan ilk Türkçe Electra NER modelidir ve 48 farklı etiket tahmini yapabilmektedir.

ÇÖZÜM NEDİR?

Modeller

- Electra-base modeli, TQUAD ve Kuzgunlar Question-Answer Dataset kullanılarak Soru-Cevap görevi için fine-tune edilmiştir. Bu model paylaşılan ilk Türkçe Electra Soru-Cevap modelidir.
- Electra-base modeli üzerine fine-tune edilmiş ilk ve tek Türkçe duygusal durum analizi modeli hazırlandı.

ÇÖZÜM NEDİR?

Kullanıcı Arayüzü

- Doğal dil işleme modellerinin hızlı ve kolay kullanımı için bir WEB API geliştirilmiştir. Ayrıca bu API'ı kullanan bir web sayfası da hazırlanmıştır. Bu sayede modeller platform ve konum bağımsız çalıştırılabilmektedir.
- Web sayfaları üzerindeki metinleri otomatik tespit edip, bu metinler üzerinde aranılan soruların cevapları bulabilen ve NER görevlerini ifa edebilen bir Chrome eklentisi geliştirilmiştir.

İŞ AKIŞI

Python Betikleri Yardımıyla Maskelenmiş Dil(MLM), NER, Soru-Cevap, Duygu Durum Analizi görevleri İçin Eğitim Verilerinin Bulunması, Düzenlenmesi ve Hazırlanması.

Veri Temini



Model Mimarisi

Kullanılacak Olan Modellerin Belirlenmesi ve Model Mimarisinin Tasarlanması. NER, Soru-Cevap, Duygu Durum Analizi Modellerinin Fine-Tune Edilmesine Karar Verilmesi.

Eğitim Betikleri

NER, Soru-Cevap, Duygu Durum Analizi Modellerinin Electra Base Modeli Üzerinden Fine-Tune Edilmesi İçin Gereken Betiklerin Hazırlanması.

Yeterli İşlem Gücü Kullanılarak Modellerin Eğitimi.

Model Eğitimi



Test

Eğitilen Modellerin Test Verileri İle Test Edilmesi ve Farklı Görevler İçin İnce Ayarların Yapılması.

İŞ AKIŞI

Modellerin Platform ve Konum
Bağımsız Kullanılabilmesi İçin Web API
Hazırlanması.

Son Kullanıcıların Modelleri
Okudukları Web Sayfalarındaki
Metinler Üzerinde Kullanabilmeleri
İçin Chrome Eklentisinin Yazılması.

Web API Hazırlanması



Modellerin Paylaşılması

Eğitimi Tamamlanan
Modellerin Huggingface
Üzerinden Paylaşılması

Web Arayüzünün Tasarımı

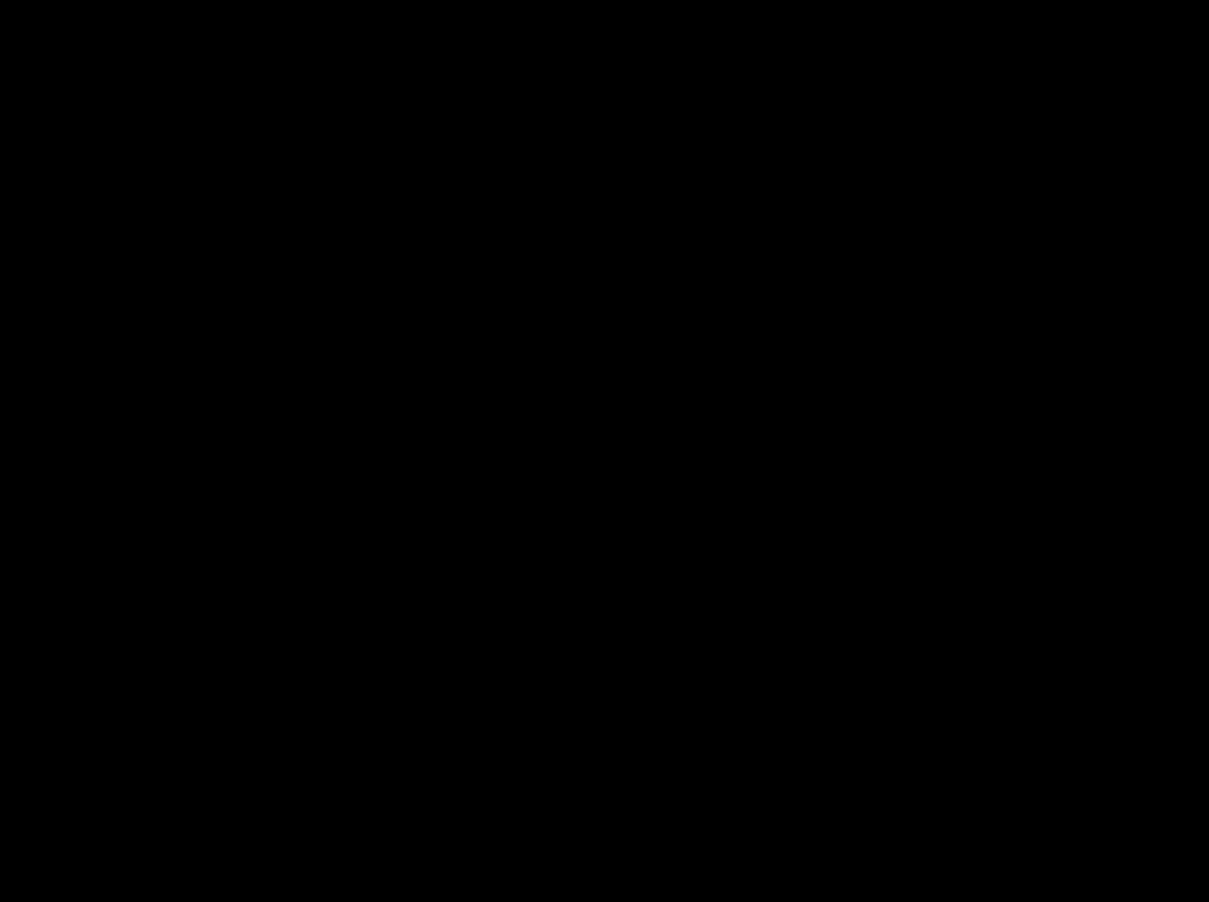
Son Kullanıcıların Modelleri
Rahatça Kullanabilmeleri İçin
API'ı Kullanan Bir Web
Sayfasının Tasarlanması.

Chrome Uzantısı



Kodların Paylaşımı

Yazılan Araçların ve Veri
Setlerinin Kullanım
Kılavuzları İle Beraber
Github Üzerinden Paylaşımı
ve Son Düzenlemelerin
Tamamlanması.



DEMO

LAB

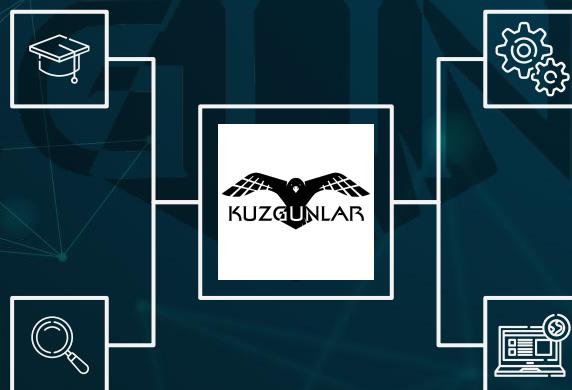
EKİP ÜYELERİNİN KATKILARI



Hüseyin ERDEM

Eğitim Betiklerinin Yazımı ve
Modellerin Eğitilmesi
Modellerin Eğitilmesi

Veri Setlerinin Çıkarımı
Modeller İçin Gereken Veri
Setlerinin Bulunması,
Düzenlenmesi ve Oluşturulması



Behçet ŞENTÜRK

Son Kullanıcının Kullanabilmesi
için Web API Yazılması
Web API Yazılması

Chrome Eklentisi
Son Kullanıcı İçin Chrome
Eklentisi Yazılması

TEŞEKKÜRLER