Türkçe Dili için Kelime Bölme Yaklaşımlarının İncelenmesi
Analysis of Subword Tokenization Approaches for Turkish Language

Erencan Erkaya, *Tunga Güngör*
Boğaziçi University, Department of Computer Engineering
Bebek, 34342 Istanbul, Turkey
{erencanerkaya@gmail.com, gungort@boun.edu.tr}

***Özetçe* —Doğal dil işleme araştırmalarında çeşitli kelime bölme yaklaşımları ortaya konulmuştur. Bu yaklaşımlar, karakter ve kelime seviyesindeki temsillerden alt kelime seviyesindeki temsillere doğru daha da gelişmiştir. Bununla birlikte, özellikle morfolojik olarak zengin diller için kelime bölme algoritmalarının model performansı üzerindeki etkisi tam olarak tartışılmamıştır. Bu çalışmada, çekimli ve morfolojik açıdan oldukça zengin bir dil olan Türkçe için kelime bölme algoritmalarının kapsamlı bir şekilde analizi yapılmıştır ve morfoloji bazlı bir yaklaşım önerilmiştir. Ayrıca, sözcük dağarcığı ve derlem boyutu gibi farklı belirteç parametrelerinin belirteçlerin özelliklerini ne şekilde değiştirdiği incelenmiştir.**

Anahtar Kelimeler—doğal dil işleme, kelime bölme algoritmaları, morfoloji, Türkçe.

***Abstract*—Various tokenization approaches have been proposed in natural language processing research. These approaches have further evolved from character- and word-level representations to subword-level representations. However, the impact of tokenizations on model performance has not been thoroughly discussed, especially for morphologically rich languages. In this paper, we comprehensively analyze subword tokenizers for Turkish which is a highly inflected and morphologically rich language and we propose a morphologically-based approach. Also, we examine how the tokenizer parameters like vocabulary and corpus sizes change the characteristics of tokenizers.**

Keywords—natural language processing, subword tokenizers, morphology, Turkish.

I. Giriş

Kelime bölme (*tokenization*), bir metni belirteç (*token*) adı verilen parçalara bölme işlemidir. Bu belirteçler sabit uzunlukta vektörler olarak temsil edilirler ve makine öğrenmesi modellerinde girdi olarak kullanılırlar. En basit kelime bölme algoritması, metni boşluk ve noktalama işaretlerine göre bölmektir. Ancak derlem boyutu arttıkça benzersiz sözcük sayısı da artacağından, bu yaklaşım bellek ve performans sorunlarına yol açmaktadır. Bir değer yaklaşım ise her kelimeyi karakterlere bölerek metni karakter dizisi olarak temsil etmektir. Bu yaklaşımda, belirteç dağarcığının boyutu benzersiz karakterlerin sayısıyla sınırlandırılmıştır. Ancak bu yöntem, kelime temsiline kıyasla çok daha fazla belirteç üreteceğinden semantik bilgi temsilini zorlaştırmaktadır. Ayrıca karakter seviyesinde belirteç kullanan yaklaşımlar, makul temsiller oluşturabilmek için daha karmaşık modellere ihtiyaç duyarlar.

Karakter ve sözcük dışında bir ara yaklaşım ise, metni alt sözcüklere (*subword*) bölmektir. Alt sözcükler kök ve ek gibi morfolojik bilgileri kodlayabilen belirteçlerden oluşur. Bu konudaki öncü algoritmalardan biri olan byte-pair encoding (BPE) [1] önce veri sıkıştırma yaklaşımı olarak kullanılmış, daha sonra alt sözcük bölme algoritması olarak uyarlanmıştır. Sinirsel dil modellerinin gelişmesiyle birlikte, alt sözcük düzeyinde girdi temsilleri popüler hale gelmiştir. Bidirectional encoder representations from transformers (BERT) [2], generative pre-training (GPT) [3] ve text-to-text transfer transformer (T5) [4] gibi modeller alt sözcük girdi temsilleri kullanırlar.

Gelişmiş modellerin çoğu kendine özel ve optimize edilmiş kelime bölme algoritmaları kullanırlar. Bu nedenle, modeller aynı kelime için farklı belirteçler üretebilmektedir. Buna ek olarak, kelime bölme algoritmaları her zaman kök, ön ek, son ek gibi anlamlı belirteçler üretmeyebilir. Belirteçlerin doğal dil işleme çalışmalarında kullanılan modellerin performansı üzerindeki etkileri, özellikle morfolojik olarak zengin diller için literatürde genel olarak tartışılmamıştır. Bu çalışmada, kelime bölme algoritmalarının morfolojiyi ne ölçüde yansıtabildiği ve anlamlı alt sözcükler üretebildiği incelenmiştir.

II. Literatür Taraması

Geçmiş yıllarda kelime bölme algoritmalarını inceleyen çeşitli çalışmalar yapılmıştır. Mielke vd. [9], kelime bölme yaklaşımlarının tarihini kapsamlı bir şekilde araştırmış ve bu algoritmaların yıllar içinde nasıl geliştiğini incelemişlerdir. Morfolojik olarak zengin diller için, her bir kelimeyi morfolojisine uygun bir şekilde temsil etmenin performansı olumlu şekilde etkilediği gösterilmiştir. Alyafeai vd. [10], morfolojik olarak zengin bir dil olan Arapça için çeşitli kelime bölme yaklaşımlarını analiz etmişlerdir. Sonuçlar kelimelerin bir bütün olarak temsil edildiği modellerin çeşitli doğal dil işleme uygulamalarında performansı arttırdığını göstermektedir.

979-8-3503-4355-7/23/$31.00 ©2023 IEEE

TABLO I: BPE, ULM ve WordPiece kelime bölme örnekleri

Kelime	BPE	ULM	Wordpiece
büyükadaya	büyü, #ka, #daya	büyük, #aday, #a	büyük, #ada, #ya
dünyalıyım	dün, #yalı, #yım	dünya, #lıyım	dünya, #lı, #yım
inceltmek	incel, #tmek	ince, #l, #t, #mek	incel, #tme, #k
kaygısını	kayg, #ısını	kaygı, #sını	kaygısı, #nı
yangının	yang, #ının	yangın, #ın	yangını, #n
modüler	mo, #düler	modül, #e, #r	modül, #er
tümüne	tüm, #üne	tümü, #n, #e	tümü, #ne
yandı	yandı	y, #andı	yan, #dı
okuru	ok, #uru	o, #kuru	okur, #u
döşek	dö, #şek	döşe, #k	döş, #ek

Domingo vd. [11], çeşitli kelime bölme algoritmalarının makine çevirisi modellerinin performansı üzerindeki etkisini incelemişlerdir. 10 dil ve 5 farklı kelime bölme yaklaşımı ile yapılan deneyler, belirteçlerin modellerin performansı üzerinde büyük etkisi olduğunu ve kelimeleri bütün olarak temsil etmenin çeviri kalitesini iyileştirdiğini göstermektedir.

Matthews vd. [12], sınırlı kelime dağarcığı boyutuna sahip modellerin aksine morfoloji bilgisinden de faydalanarak karakter tabanlı bir model geliştirmiştir. Sonuçlar, önerilen modelin, n-gram modelinden daha iyi çalıştığını göstermektedir. Özellikle Türkçe gibi morfolojisi zengin diller için, morfoloji bilgisinin model içerisinde kullanılması performansı iyileştirmektedir.

III. Materyal ve Metotlar

Kelimelerin alt sözcüklere ayrılması amacıyla literatürde en yaygın olarak kullanılan yöntemler byte-pair encoding (BPE) [1], unigram language model (ULM) [5] ve WordPiece [6] algoritmalarıdır. BPE algoritması her iterasyonda derlem içerisinde yer alan frekansı en yüksek olan belirteç çiftini birleştirir. Bu süreç önceden belirlenmiş olan kelime dağarcığı boyutuna ulaşılana kadar devam eder. WordPiece algoritması da BPE algoritmasına benzer bir yaklaşımı takip eder. İki algoritma arasındaki temel fark, WordPiece algoritmasının her iterasyonda çiftlerin frekansı yerine bir dil modelinden faydalanarak yeni belirteçleri oluşturmasıdır. ULM algoritması ise hedef kelime dağarcığı boyutundan çok daha büyük olan bir belirteç listesi ile başlar. Hedef kelime dağarcığı boyutuna ulaşıncaya dek bir dil modelinden yararlanarak kademeli bir şekilde kelime listesinden belirteçleri çıkarır. Tablo I'de bahsedilen üç yöntemle ilgili kelime bölme örnekleri gösterilmiştir.

Belirtilen algoritmalar dilden bağımsız olarak çalıştıkları için bir dilin morfoloji bilgisini kullanmazlar. Bu nedenle, kelimeler çoğunlukla ek ve köklerine uygun bir şekilde temsil edilmezler (Tablo I). Bu çalışmada, morfolojiye uyumluluğun modeller üzerindeki etkisini ölçebilmek amacıyla *WordPiece Morphology* olarak adlandırılan bir yöntem geliştirilmiştir. Bu yöntem WordPiece algoritmasının girdi temsilleri üzerinde değişiklikler yapılarak elde edilmiştir. Eğitim aşamasında, ilk olarak kelimeler morfolojik analizden geçirilerek kök ve ekler belirlenir. Ekler özel karakterler arasına alınarak bölünemez semboller olarak temsil edilirler ve belirteç birleştirme sürecine dahil edilmezler. Örneğin, *kalemleri* kelimesi *kalem<ler><i>* olarak temsil edilir. Bölme algoritması tarafından kelime içindeki *<ler>* ve *<i>* kısımlarının ek olduğu anlaşılmış olur. Eğitim sonunda, ekler ve kökler farklı belirteçler kullanılarak temsil edilmiş olurlar. Bu yöntemde bir morfoloji analiz aracı kullanılmasına rağmen, kelime bölme işleminin sonucu morfolojik bölmeden farklıdır ve kelimeler alt sözcüklere bölünürler.

Kelime bölme algoritmaları aynı kelime için farklı belirteçler üretebilmektedirler. Algoritmaların çeşitli özelliklerini ve dilin morfolojisini ne ölçüde yansıtabildiklerini ölçebilmek için bu çalışmada aşağıdaki metrikler tanımlanmıştır.

- Tek Kelime Belirteçler: Bir derlemde tek bir belirteç kullanılarak ifade edilen sözcüklerin yüzdesini belirtir. Örneğin, *kalemler* kelimesi *kalemler* olarak bölünmüşse, tek kelimelik bir belirteç olarak görülür. Bu metrik, bir belirteç oluşturucunun, sözcükleri alt sözcüklere bölmeden bir derlemi ne kadar iyi temsil ettiğini gösterir.
- Morfoloji Uyumlu Belirteçler: Morfolojik bölümleme ile aynı şekilde bölünmüş kelimelerin yüzdesini ölçer. Örneğin, *kalemler* kelimesi *kalem* ve *ler* olarak bölünmüşse, morfoloji ile uyumlu bir bölünmedir. Bu metriğin yüksek olması, kelime bölümleme algoritmasının morfolojiyi daha iyi temsil ettiğini gösterir.
- Üretkenlik: Kelime başına ortalama belirteç sayısını belirtir. Belirteç modelinin kelimeleri ne sıklıkta kelimelere böldüğünü gösterir. Üretkenliği yüksek bir model, daha kısa belirteçler üretme eğilimindedir.
- Ortalama Belirteç Uzunluğu: Bir kelime dağarcığındaki belirteç başına ortalama karakter sayısını ölçer.

IV. Deneyler

Kelime bölme algoritmalarının her biri aynı parametreler kullanılarak Open Super-large Crawled Aggregated coRpus (OSCAR) [7] derlemi üzerinde eğitilmiştir. Daha önce ifade edildiği üzere, aynı kelime için algoritmalar farklı belirteçler üretebilmektedir ve üretilen belirteçler her zaman anlamlı kök ve eklerden oluşmayabilir. Kelime bölme algoritmalarını bu açılardan karşılaştırabilmek için test veri seti olarak BOUN Treebank derlemi [8] kullanılmıştır. Derlem en az bir eke sahip 44275 kelime ve toplamda 97424 kelime içermektedir. Derlemdeki tüm kelimeler eğitilen belirteç modelleri kullanılarak alt sözcüklere bölünmüş ve Bölüm III'de açıklanan metrikler kullanılarak analiz edilmiştir. Sonuçlar Tablo II'de verilmiştir. BPE diğer algoritmalara kıyasla daha çok tek kelime belirteç üretmektedir. WordPiece algoritması ise ek ve kök bütünlüğü açısından BPE ve ULM algoritmlarına göre daha iyi performans göstermektedir. WordPiece Morphology modeli ise kelime başına daha fazla belirteç üretmekte ve Türkçenin morfolojisini daha iyi kodlayabilmektedir.

TABLO II: BPE, ULM, WordPiece ve WordPiece Morphology algoritmalarının karşılaştırılması

	BPE	ULM	WordPiece	WordPiece Morphology
Tek Kelime Belirteçler	%75	%68	%73	%47
Morfoloji Uyumlu Belirteçler	%52	%51	%54	%83
Üretkenlik	1.31	1.50	1.34	1.99
Üretkenlik (Ek almış kelimeler)	1.54	1.78	1.58	2.95
Ortalama Belirteç Uzunluğu	6.10	4.43	5.72	5.00

Kelime dağarcığı ve derlem boyutlarının kelime bölme algoritmaları üzerindeki etkisini görmek için de deneyler yapılmıştır. Bu amaçla, algoritmalar farklı kelime dağarcığı ve derlem boyutları ile OSCAR derleminde eğitilmiş ve BOUN Treebank derleminde test edilmiştir. Kelime dağarcığı boyutu sonuçları Tablo III'de ve Şekil 1'de gösterilmektedir. Tablo III'de örnek olarak ULM algoritmasının iki metriğe göre sonuçları verilmiştir. Kelime dağarcığı boyutu arttıkça, kelime başına daha az belirteç üretilmektedir. 1000 kelime dağarcığı boyutunda eğitilmiş model, her bir karakter için bir belirteç üretmektedir. Kelime dağarcığı boyutu derlem içerisindeki benzersiz kelime sayısına yakın olduğunda ise, modelin üretkenliği 1'e doğru azalmaktadır. Eki olan kelimelerin genellikle eki olmayan kelimelerden daha uzun olması sebebiyle, eki olan kelimeler daha fazla belirteç kullanılarak temsil edilmektedir.

TABLO III: Kelime dağarcığı boyutunun üretkenlik ve belirteç uzunluğuna etkisi

Kelime Dağarcığı Boyutu	Üretkenlik	Üretkenlik (Ek almış kelimeler)	Ortalama Belirteç Uzunluğu
1000	5.47	8.03	1.01
2000	5.35	7.87	1.05
3000	3.08	4.21	1.17
5000	2.26	3.01	1.93
10000	2.02	2.61	2.24
20000	1.73	2.16	2.86
30000	1.52	1.81	4.23

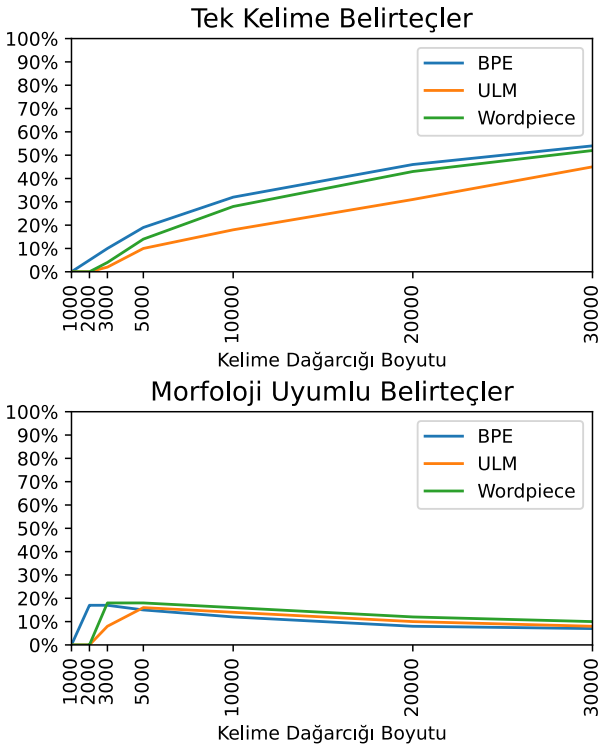
Şekil 1'de ise diğer iki metriğe göre BPE, WordPiece ve ULM yöntemlerinin sonuçları örnek olarak ek almış kelimeler üzerinde karşılaştırılmıştır. Kelime dağarcığı boyutu arttıkça, tek kelime belirteçlerin sayısı artmaktadır. Öte yandan, morfoloji uyumlu belirteçlerin sayısı özellikle 5000 kelime dağarcığı boyutundan sonra önemli ölçüde değişmemektedir. Buna göre, 3000 ve 5000 kelime dağarcığı boyutlarının Türkçenin morfolojisini iyi bir şekilde temsil edebildiği görülmektedir.

Kelime dağarcığı boyutunun etkisi örnek bir kelime üzerinde (*hazinesidir*) Tablo IV'de gösterilmiştir. Kelime, farklı kelime dağarcığı boyutları ile eğitilmiş WordPiece modellerine göre belirteçlere ayrılmıştır. 1000 kelime dağarcığı boyutunda eğitilmiş model, her bir karakter için bir belirteç üretmektedir. 30000 kelime dağarcığı ile eğitilmiş model ise kelimeyi tam olarak morfolojisine uygun bir şekilde temsil etmektedir.

TABLO IV: *hazinesidir* kelimesinin farklı kelime dağarcığı boyutlarına göre bölünmesi

Kelime Dağarcığı Boyutu	Belirteçler
1000	h, #a ,#z ,#i ,#n ,#e ,#s ,#i ,#d, #i, #r
2000	h, #a ,#z ,#i ,#n ,#e ,#s ,#i ,#d, #i, #r
3000	haz, #ine ,#si ,#dir
5000	haz, #ine ,#si ,#dir
10000	haz, #ine ,#si ,#dir
20000	hazin, #esi ,#dir
30000	hazine, #si ,#dir

Derlem boyutunun modeller üzerindeki etkisi ise Tablo V'de ve Şekil 2'de gösterilmiştir. Kelime dağarcığı deneylerine benzer şekilde, Tablo V'de ULM algoritmasının sonuçları, Şekil 2'de de ek almış kelimeler üzerindeki sonuçlar verilmiştir. Tablodaki sonuçlara göre, derlem boyutu artıkça ortalama belirteç uzunluğu azalmakta ve kelime başına ortalama belirteç



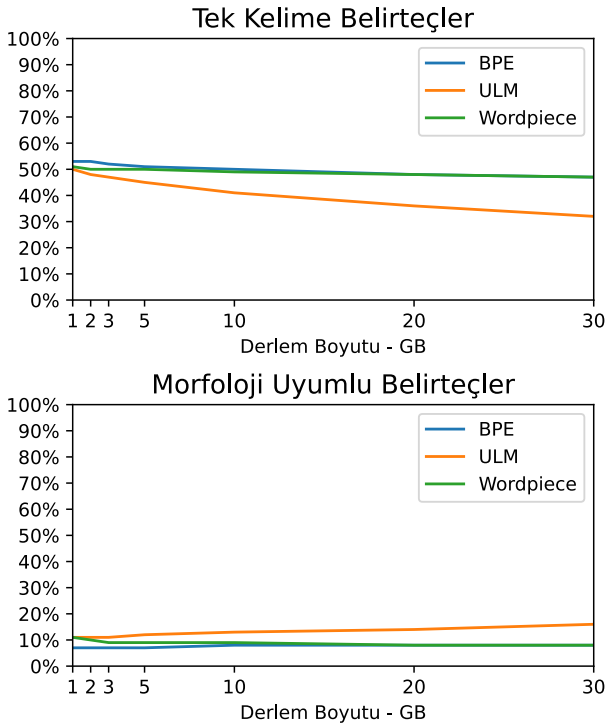
Şekil 1: Kelime dağarcığı boyutunun tek kelimelik belirteç ve morfoloji uyumlu belirteçlere etkisi

TABLO V: Derlem boyutunun üretkenlik ve belirteç uzunluğuna etkisi

Derlem Boyutu - GB	Üretkenlik	Üretkenlik, yalnızca son eke sahip kelimeler	Ortalama Belirteç Uzunluğu
1	1.37	1.61	6.07
2	1.39	1.63	5.83
3	1.40	1.65	5.68
5	1.42	1.67	5.43
10	1.44	1.70	5.09
20	1.47	1.74	4.75
30	1.49	1.77	4.50

sayısı artmaktadır. Şekildeki sonuçlara göre de, derlem boyutu arttıkça tek kelimelik belirteçlerin sayısı azalmaktadır. Bunun yanı sıra, morfoloji uyumlu belirteçlerin sayısı artmaktadır. Bu durum, algoritmaların daha fazla veri ile eğitildiğinde modellerin dilin morfolojisini daha iyi temsil ettiğini göstermektedir. ULM algoritması diğer algoritmalar arasında derlem boyutuna karşı en duyarlı olan algoritmadır. Derlem boyutunun artışı tek kelimelik belirteçlerin sayısında bir azalmaya neden olur, ancak BPE ve WordPiece algoritmalarına göre morfoloji uyumlu belirteçlerin sayısını büyük ölçüde arttırır.

Kelime bölme algoritmalarının Bölüm III'de açıklanan metriklere göre yapılan içsel değerlendirmesine (*intrinsic evaluation*) ek olarak, morfoloji bazlı kelime bölme yönteminin iki farklı doğal dil işleme uygulaması üzerinde dışsal değerlendirmesi de (*extrinsic evaluation*) yapılmıştır. Bu amaçla, OSCAR derlemi üzerinde ELECTRA [13] modeli kullanılarak iki farklı modelin ön eğitimi (*pretrain*) yapılmıştır. Modellerden biri WordPiece ile üretilen belirteçleri, diğeri WordPiece Morphology ile üretilen belirteçleri kullanarak eğitilmiştir.



Şekil 2: Derlem boyutunun tek kelimelik belirteç ve morfoloji uyumlu belirteçlere etkisi

TABLO VI: Duygu analizi uygulaması sonuçları

	WordPiece	WordPiece Morphology
Accuracy	0.86	**0.88**

İlk uygulama duygu analizi uygulamasıdır. ELECTRA modelinin üzerine [CLS] vektörünü girdi olarak alan bir sınıflandırma katmanı eklenmiştir ve model 5331 pozitif ve 5331 negatif yorumdan oluşan BeyazPerde veri seti [14] ile eğitilmiştir. Modellerin doğruluk sonuçları Tablo VI'da verilmiştir. İkinci uygulama olan soru cevaplama uygulaması için Hugging Face soru cevaplama kütüphanesi ve 9200 soru-cevap çiftinden oluşan TQuAD veriseti [15] kullanılmıştır. Model, cevabın başlangıç ve bitiş indekslerini tahmin etmeye yönelik eğitilmiştir. Modellerin F1 ve EM (*exact match*) sonuçları Tablo VII'de verilmiştir. Her iki uygulamada da, morfoloji bazlı kelime bölme modelinin performansı arttırdığı görülmektedir.

TABLO VII: Soru cevaplama uygulaması sonuçları

	WordPiece	WordPiece Morphology
F1	57.38	**58.09**
EM	37.78	**38.17**

V. Sonuç

Bu çalışmada, en yaygın kullanılan kelime bölme algoritmalarından BPE, WordPiece ve ULM analiz edilmiştir. Bunun yanı sıra, morfoloji açısından optimize edilmiş bir kelime bölme algoritması önerilmiştir. Algoritmaların Türkçe gibi morfolojisi zengin bir dili ne ölçüde kodlayabildiğini değerlendirmek için çeşitli metrikler tanımlanmıştır. Aynı derlem ve parametrelerle eğitilmiş modeller, tüm kelimelerin yaklaşık olarak yarısını tek kelime belirteçler kullanarak temsil etmektedir. Kelime dağarcığı boyutunun ve derlem boyutunun algoritmaların performansını nasıl etkilediği incelenmiştir. Kelime dağarcığı boyutunun, kelimelerin morfolojiye uygun olarak bölünmesinde önemli bir etkisi bulunmaktadır. Ayrıca, morfoloji bazlı kelime bölme algoritmasının çeşitli doğal dil işleme uygulamalarında performansı iyileştirdiği gösterilmiştir.

Kaynaklar

[1] Gage, P., "A New Algorithm for Data Compression", C Users Journal, Vol. 12, No. 2, pp. 23–38, 1994.

[2] Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 4171–4186, ACM, Minnesota, 2019.

[3] Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, Vol. 33, pp. 1877–1901, 2020.

[4] Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-toText Transformer", Journal of Machine Learning Research, Vol. 21, No. 140, pp. 1–67, 2020.

[5] Kudo, T., "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates", Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 66–75, ACL, Melbourne, 2018.

[6] Schuster, M. and K. Nakajima, "Japanese and Korean Voice Search", International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5149–5152, 2012.

[7] Abadji, J., P. J. O. Suarez, L. Romary and B. Sagot, "Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus", Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9), pp. 1–9, Mannheim, 2021.

[8] Türk, U., F. Atmaca, S¸. B. Özateş, G. Berk, S. T. Bedir, A. Köksal, B. Ö. Başaran, T. Güngör and A. Özgür, "Resources for Turkish Dependency Parsing: Introducing the BOUN Treebank and the BoAT Annotation Tool", Language Resources and Evaluation, pp. 1–49, 2021.

[9] Mielke, S., Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Galle, A. Raja, C. Si, W. Lee, B. Sagot and S. Tan, "Between Words and Characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP", arXiv preprint arXiv:2112.10508 , 2021.

[10] Alyafeai, Z., M. S. Al-shaibani, M. Ghaleb and I. Ahmad, "Evaluating Various Tokenizers for Arabic Text Classification", arXiv preprint arXiv:2106.07540 , 2021.

[11] Domingo, M., M. Garcıa-Martınez, A. Helle, F. Casacuberta and M. Herranz, "How Much Does Tokenization Affect Neural Machine Translation?", arXiv preprint arXiv:1812.08621 , 2018.

[12] Matthews, A., G. Neubig and C. Dyer, "Using Morphological Knowledge in Openvocabulary Neural Language Models", Proceedings of the Conference of the North American Chapter of the ACL: Human Language Technologies, Vol. 1, pp. 1435–1445, 2018.

[13] Clark, K., M.-T. Luong, Q. V. Le and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", International Conference on Learning Representations (ICLR), 2020.

[14] Demirtaş, E. and M. Pechenizkiy, "Cross-lingual Polarity Detection with Machine Translation", Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, pp. 1–8, 2013.

[15] Soygazi, F., O. Çiftçi, U. Kök and S. Cengiz, "THQuAD: Turkish Historic Question Answering Dataset for Reading Comprehension", Conference on Computer Science and Engineering, pp. 215–220, 2021.