Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd
https://doi.org/10.53070/bbd.990959

Journal of Computer Science
ISSN: 2548-1304
Volume: IDAP-2021, Issue: Special, pp: 335-340
@Anatolian Science

Türkçe-İngilizce Akademik Çeviriler için Paralel Corpora Oluşturulması

Creating a Parallel Corpora for Turkish-English Academic Translations

İlhami SEL[*1], Hüseyin ÜZEN[2], Davut HANBAY[1]

[1] Bilgisayar Mühendisliği Bölümü, İnönü Üniversitesi, Malatya, Türkiye

[2]Bilgisayar Mühendisliği Bölümü, Bingöl Üniversitesi, Bingöl, Türkiye

(ilhamisel23@gmail.com, huzen@bingol.edu.tr, davut.hanbay@inonu.edu.tr)

Received: Sep.3, 2021 Accepted: Sep.16, 2021 Published: Oct.20, 2021

Özetçe— Paralel corpora aynı anlama gelen cümlelerin farklı dillerde temsil edilmesiyle oluşturulan veri setleridir. Makine çeviri sistemlerinde kaliteyi belirleyen en önemli öğelerden birisi büyük miktarda ve yüksek kalitede oluşturulmuş paralel corporadır. Türkçe – İngilizce dil çifti için oluşturulan bu tür veriler genellikle yetersizdir. Bu çalışmada Türkçe – İngilizce dilleri arasında akademik çeviriler için kullanılabilecek büyük miktarda paralel corpora oluşturulmuştur. Bu veri seti oluşturulurken lisansüstü tezlerinin özet kısımları kullanılmıştır. Vecalign ve Hunalign gibi cümle hizalama algoritmaları kullanılarak en iyi eşleştirmeler elde edilmiştir. Yapılan çalışmalar sonucunda 1M paralel cümle çifti elde edilmiştir. Ayrıca elde edilen verinin kalitesini ölçebilmek için Bi-LSTM tabanlı çeviri sistemi oluşturulmuştur. Oluşturulan model TED(Tr-En) test seti üzerinde sıfır vuruş öğrenme (zero shot learning) yöntemiyle 15.8 Bleu puanı elde etmiştir.

Anahtar Kelimeler: *Paralel Corpora*, *Sinirsel Makine Çevirisi*, *Cümle Hizalama*, *Doğal Dil İşleme.*

Abstract— Parallel corpora are data sets created by representing sentences with the same meaning in different languages. One of the most important elements that determine the quality in machine translation systems is the parallel corpora created in large quantities and with high quality. Such data for the Turkish – English language pair are generally insufficient. In this study, a large amount of parallel corpora has been created that can be used for academic translations between Turkish and English languages. While creating this data set, the abstracts of the postgraduate theses were used. The best matches were obtained using sentence alignment algorithms such as Vecalign and Hunalign. As a result of the studies, 1M parallel sentence pairs were obtained. In addition, an Bi-LSTM-based translation system was created to measure the quality of the obtained data. The created model obtained 15.8 Bleu points with zero-shot learning method on the TED (Tr-En) test set.

Keywords: *Parallel Corpora*, *Neural Machine Translation*, *Sentence Alignment*, *Natural Language Processing.*

1. Giriş

Derin sinir ağları ve Doğal dil işleme alanındaki gelişmelere bağlı olarak makine çevirisi oldukça popüler hale gelmiştir. Makine çevirisi akademik öneminin yanında büyük teknoloji şirketlerinin de

(Google[1], Yandex[2], Bing[3]) yatırım yaptığı alanlardan birisidir. Ayrıca son yıllarda Transformer (Vaswani vd. 2017) ağlarına bağlı olarak yüksek çevirim kalitesi elde etmişlerdir. Tüm bu gelişmelerin yanında çevirim kalitesini etkileyen en önemli unsur büyük miktarda ve yüksek kalitede hazırlanmış paralel corporadır (Barrault vd. 2019). İngilizce-Almanca veya İngilizce-Fransızca gibi dillere ait makine çevirileri düşük kaynaklı dillere oranla daha başarılı sonuçlar elde etmişlerdir (Bawden vd. 2020). Bu çalışma kapsamında Türkçe İngilizce makine çeviri kalitesini artırabilmek için paralel corpora oluşturulmaya çalışılmıştır.

1.1. İlgili Çalışmalar:

Makine çeviri sistemlerinde Türkçe dili düşük kaynaklı (low resources) diller arasında gösterilmektedir(Chaudhary vd. 2019; Ataman 2018) ve genellikle çalışmalar veri setini artırmaya yöneliktir. 2018 yılında Ataman ve ark. tarafından Bianet haber sitesinden alınan makaleler kullanılarak Türkçe, Kürtçe ve İngilizce paralel veri seti oluşturulmuştur (Ataman 2018). Bu veri seti Tr-En dilleri arasında 35k adet cümleden oluşmuştur. Ted konuşma videolarının alt yazıları kullanılarak oluşturulan paralel veri setinde (Qi vd. 2018) ise 191k adet paralel cümle bulunmaktadır. 2019 yılında Wikipedia makalelerinin kullanıldığı çalışmada ise (Schwenk vd. 2021) 477k adet paralel cümleden oluşan veri seti sunulmuştur. Bu alanda ki en kapsamlı çalışma Facebook AI takımı tarafından yapılmıştır. 392M web sayfasının tarandığı çalışma da Tr-En dahil 8144 dil çifti için veri seti oluşturulmuştur (El-Kishky vd. 2020).

2. Materyal ve Metod

2.1. Materyal:

Türkçe İngilizce paralel corpora için lisansüstü tezlerden faydalanılmıştır. Türkiye'de tezlerin özet kısımları Türkçe ve İngilizce olmak üzere iki farklı dilde hazırlanmaktadır ve Yöktez[4] (Ulusal Tez Merkezi) sistemi bu tezleri yayınlamaktadır. Veri madenciliği yöntemleri ile tezler alanlarına göre hazırlanan bot tarafından taranmış, "Özet" ve "Abstract" kısımları ayrı ayrı kaydedilmiştir. 178 farklı alandan 2015-2020 yılları arasında yayınlanmış 238,233 tez bu çalışma için kullanılmıştır. İlk 10 alan ve veri setindeki dağılım Tablo-1'de gösterilmiştir.

Tablo 1: Taranan tezler ve alanlara göre dağılımı

Alanlar	**Oran (%)**
Eğitim ve Öğretim	6.1
İşletme	3.1
Psikoloji	2.9
Ziraat	2.77
Makine Mühendisliği	2.73
Kimya	2.45
Tarih	2.44
Türk Dili ve Edebiyatı	2.29
Matematik	2.23
Bilgisayar Bilimleri	2.18
Diğerleri	68.5

2.2. Metin Önişleme:

Metin ön işleme adımları aşağıdaki gibidir:

[1] https://translate.google.com/
[2] https://translate.yandex.com/
[3] https://www.bing.com/translator
[4] https://tez.yok.gov.tr/UlusalTezMerkezi/

- Sırayla her dosya okunarak bir corpora oluşturulmuştur. Bu adımda birden fazla alan etiketi ile yayınlanan aynı tezler filtrelenmiştir.
- Veriler cümlelere bölünmüştür. Veriler bu aşamada yaklaşık 3M Türkçe ve İngilizce cümleden oluşmaktadır.
- Metin küçük harflere çevrilmiştir. Metinden özel karakterler ve noktalama işaretleri ve sayılar çıkarılarak sadece harflerden oluşması sağlanmıştır.
- Boyutları 400 karakterden büyük ve 20 karakterden küçük olan cümleler yok sayılmıştır.

2.3. Cümle Hizalama:

Bu çalışma kapsamında (Haddow ve Kirefu 2020) literatürde kullanıldığı şekilde iki farklı cümle hizalama algoritması Hunalign (Varga vd. 2005), Vecalign (Thompson ve Koehn 2020) kullanılmıştır. Sonrasında bu iki algoritmanın sonuçları birleştirilerek kesişen cümleler seçilmiştir.

2.3.1. Hunalign Cümle Hizalama Algoritması:

Hunalign (Varga vd. 2005) iki dilli metni cümle düzeyinde hizalar. Hunalign bir sözlük varlığında onu kullanır ve bu bilgiyi Gale-Church cümle uzunluğu bilgisiyle birleştirir. Sözlük olmadığında önce cümle uzunluğu bilgisiyle hizalar sonrasında bu hizalamaya bağlı olarak sözlük oluşturur. Ardından oluşturulan sözlüğü kullanarak metni ikinci geçişte yeniden hizalar. Bu çalışma da Hunalign algoritmasında Pavlick ve ark. (Pavlick vd. 2014) tarafından hazırlanan sözlük kullanılmıştır.

2.3.2. Vecalign Cümle Hizalama Algoritması:

Vecalign (Thompson ve Koehn 2020) iki dilli cümle yerleştirmelerinin (embedding) benzerliğine dayanan yeni bir cümle hizalama puan işlevini kullanarak cümleleri hizalamaya çalışmaktadır. Algoritma cümle yerleştirmesi için ön eğitimli LASER çok dilli (93 dilde) cümle yerleştirme (Artetxe ve Schwenk 2019) modelini kullanmaktadır. Vecalign cümle hizalaması iki bölümde incelenebilir:

1. Bir veya daha fazla bitişik kaynak cümleyi ve bir veya daha fazla bitişik hedef cümleyi alan ve birbirlerinin çevirisi olma olasılığını gösteren bir puan döndüren bir puan işlevi,
2. Yukarıdaki puan işlevini kullanarak iki belgeyi alan ve bir hipotez hizalaması döndüren bir hizalama algoritması.

Çok dilli cümle yerleştirmeleri arasındaki normalleştirilmiş kosinüs mesafesine dayalı yeni bir puanlama işlevi sunarak (1'de), dinamik programlama yaklaşımının yeni bir uygulamasıyla bağlantılı olarak uygulanmaktadır (2'de) (Thompson ve Koehn 2020).

Kosinüs benzerliği (denklem-1) (Sel vd. 2019; Minaee vd. 2020) cümle vektörlerini karşılaştırmak için sık kullanılan bir yöntemdi.

$$\cos(\theta) = benzerlik(A,B) = \frac{A.B}{||A|| \times ||B||} \quad (1)$$

Cümle hizalaması minimum paralel birimler arar, ancak kosinüs benzerliğine sahip Dinamik Programlama çoktan çoğa hizalamaları desteklemektedir. Örneğin, üç 1–1 hizalama bildirmesi gerektiğinde 3–3 hizalamasının sağlanması (Thompson ve Koehn 2020). Bu sorunu çözmek için maliyeti, belirli bir hizalamada dikkate alınan kaynak ve hedef cümlelerin sayısına göre ölçeklendirilmiştir. Ortaya çıkan puanlama maliyeti fonksiyonu denklem-2 de verilmiştir.

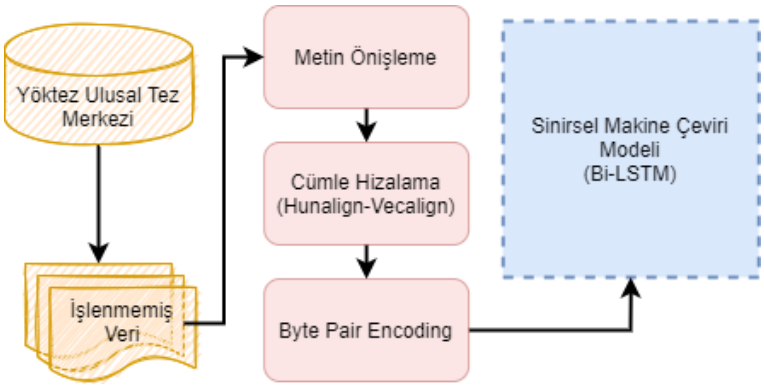
$$c(x,y) = \frac{(1-\cos(x,y))nSents(x)nSents(y)}{\sum_{s=1}^{S} 1-\cos(x,y_s) + \sum_{s=1}^{S} 1-\cos(x_s,y)} \quad (2)$$

Denklem de x, y kaynak-hedef belgeden bir veya daha fazla ardışık cümleyi belirtir, cos(x,y) yerleştirme vektörleri arasındaki benzerliği göstermektedir. nSents(x), nSents(y) x, y'deki cümle sayısını belirtmektedir ve $x_1,...,x_S$ ile $y_1,...,y_S$ verilen belgeden düzgün olarak alınan örneklerdir (Thompson ve Koehn 2020).

Cümle hizalama algoritmaları sonucunda kesişen cümleler seçilmiştir. Sonuç olarak 1.062.588 paralel cümleden oluşan corpora elde edilmiştir.

3. Sinirsel Makine Çeviri Uygulaması

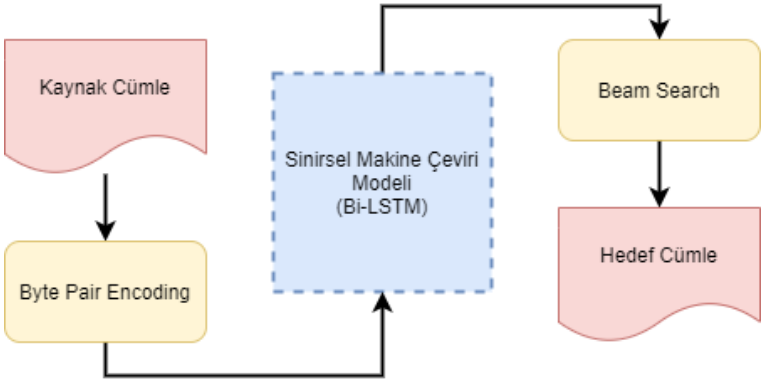
Oluşturulan veri setinin başarısını test edebilmek için Bi-LSTM (Yang vd. 2020) tabanlı temel dikkat mekanizmasına ve Encoder-Decoder mimarisine sahip sinirsel makine çeviri sistemi tasarlanmıştır (Şekil 1).



Şekil 1: Oluşturulan makine çeviri sistemi

Uygulama şu işlemlerden oluşur:

- Yöktez sisteminden işlenmemiş veri (Raw Data) veri madenciliği yöntemleri ile text dosyalarına kaydedilmiştir.
- Ham veri metin önişleme adımlarından geçirilerek noktalama işaretleri, anlamsız işaretler, rakamlar silinerek metin küçük harflere dönüştürülmüştür.
- Cümle hizalama algoritmalarından geçirilerek paralel corpora oluşturulmuştur.
- Corpora da az sayıda geçen veya hiç geçmeyen kelime sorunlarını gidermek için kelimeler parçalara ayrılarak byte çifti şeklinde kodlanmıştır (Sennrich vd. 2016).
- 2 katmana sahip ve tam ileri beslemeli ağ sayısı 1024 (N=2, FFNN=1024) olan Bi-LSTM çeviri modeli oluşturulmuştur. Modelin parametre sayısı yaklaşık 31M olmuştur.
- Model bulut hesaplama yöntemiyle 16Gb hafızaya sahip ekran kartı üzerinde eğitilmiştir.
- Tüm işlemler için python programlama dili kullanılmıştır. Veri madenciliği için ‘Selenium’, metin önişleme için ‘NLTK’ makine çeviri sistemi için ‘Pytorch-Fairseq’ kütüphaneleri kullanılmıştır.
- Model eğitildikten sonra çeviri kalitesini artırabilmek için Işın arama (Beam Search) (Şekil-2) yöntemi kullanılmıştır (Britz vd. 2017).



Şekil 2: Cümle çeviri sistemi

- Model sadece oluşturulan paralel corpora üzerinde eğitilmiştir. TED (Qi vd. 2018) eğitim seti modelin eğitimi esnasında kullanılmamıştır.

- Eğitilen model TED test seti üzerinde iki yönlü çeviri yaparak başarımı Bleu (Post 2018) (denklem 3) puanı ölçülerek hesaplanmıştır.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count_{clip}(n-gram')} \quad (3)$$

- Bu yöntemle model eğitim seti üzerinde eğitilmeden sadece test seti üzerinde çeviri yapılması sıfır vuruşlu eğitim (Zero Shot Learning) yöntemi olarak adlandırılmaktadır (Johnson vd. 2017).
- Oluşturulan paralel corpora ve sinirsel makine çeviri sistemi TED test seti üzerinde Türkçe İngilizce çeviriler için 15.8, İngilizce Türkçe çeviriler için 14.9 Bleu puanı elde etmiştir.

4. Sonuç

Yapılan çalışma da Türkçe İngilizce dilleri arasında çeviri sistemi oluşturabilmek için 1M cümleye sahip paralel corpora oluşturulmuştur. Oluşturulan paralel corpora literatürde ulaşabildiğimiz tüm veri setlerinden daha büyük miktardadır. Paralel corpora'yı test edebilmek için oluşturulan Sinirsel Makine Çeviri sistemi farklı bir veri setinde ≈15 Bleu puanı alarak başarılı bir sonuç elde etmiştir. Sonraki çalışmalarda İngilizce makale yazımı için hem bir çeviri sistemi hem de "proofreading" olarak adlandırılan çevirilerdeki hataların tespiti için uygulamalar geliştirilmeye çalışılacaktır. Ayrıca tüm veri ve kodlara Github[5] sayfamızdan ulaşılabilir.

Kaynaklar

Artetxe, Mikel, and Holger Schwenk. 2019. "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond." *Transactions of the Association for Computational Linguistics* 7: 597–610. https://doi.org/10.1162/tacl_a_00288.

Ataman, Duygu. 2018. "Bianet: A Parallel News Corpus in Turkish, Kurdish and English," 1–4. http://arxiv.org/abs/1805.05095.

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, et al. 2019. "Findings of the 2019 Conference on Machine Translation (WMT19)" 2 (Day 1): 1–61. https://doi.org/10.18653/v1/w19-5301.

Bawden, Rachel, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, et al. 2020. "Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages." *Proceedings of the Fifth Conference on Machine Translation*, 660–87. https://www.aclweb.org/anthology/2020.wmt-1.76.

Britz, Denny, Anna Goldie, Minh Thang Luong, and Quoc V. Le. 2017. "Massive Exploration of Neural Machine Translation Architectures." *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1442–51. https://doi.org/10.18653/v1/d17-1151.

Chaudhary, Vishrav, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. "Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings" 3 (Day 2): 261–66. https://doi.org/10.18653/v1/w19-5435.

El-Kishky, Ahmed, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. "CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs," 5960–69. https://doi.org/10.18653/v1/2020.emnlp-main.480.

Haddow, Barry, and Faheem Kirefu. 2020. "PMIndia -- A Collection of Parallel Corpora of Languages of India." http://arxiv.org/abs/2001.09907.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, et al. 2017. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *Transactions of the Association for Computational Linguistics* 5: 339–51. https://doi.org/10.1162/tacl_a_00065.

Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. "Deep Learning Based Text Classification: A Comprehensive Review." *ArXiv* 54 (3).

Pavlick, Ellie, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. "The

[5] https://github.com/ilhamisel/SentenceAlignment

Language Demographics of Amazon Mechanical Turk." *Transactions of the Association for Computational Linguistics* 2: 79–92. https://doi.org/10.1162/tacl_a_00167.
Post, Matt. 2018. "A Call for Clarity in Reporting BLEU Scores." *Proceedings of the Third Conference on Machine Translation: Research Papers*, April, 186–91. https://doi.org/10.18653/v1/W18-6319.
Qi, Ye, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. "When and Why Are Pre-Trainedword Embeddings Useful for Neural Machine Translation?" *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 2: 529–35.
Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia." *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 1351–61.
Sel, İlhami, Ali Karci, and Davut Hanbay. 2019. "Karşılıklı Bilgi Kullanılarak Metin Sınıflandırma İçin Özellik Seçimi Feature Selection for Text Classification Using Mutual Information." *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 18–21.
Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. "Neural Machine Translation of Rare Words with Subword Units." *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers* 3: 1715–25.
Thompson, Brian, and Philipp Koehn. 2020. "Vecalign: Improved Sentence Alignment in Linear Time and Space." *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 1342–48. https://doi.org/10.18653/v1/d19-1136.
Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. "Parallel Corpora for Medium Density Languages." *International Conference Recent Advances in Natural Language Processing, RANLP* 2005-Janua (2003): 590–96. https://doi.org/10.1075/cilt.292.32var.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 2017-Decem (Nips): 5999–6009.
Yang, Shuoheng, Yuxin Wang, and Xiaowen Chu. 2020. "A Survey of Deep Learning Techniques for Neural Machine Translation." http://arxiv.org/abs/2002.07526.