BÖLÜM 7

DOĞAL DİL İŞLEME UYGULAMALARI VE YAKLAŞIMLARI

Züleyha YİNER[1]

GİRİŞ

İnsanların en temel iletişim araçları doğal dillerdir. Bilişim teknolojilerinin gelişmesi ile yeni ilgi alanları bu doğal dillerin bilgisayarlar tarafından tanınmasını sağlaması yönünde olmuştur ve bu alandaki çalışmalar bilişim teknolojileri için vazgeçilmez olmuştur. Bilgisayar bilimi ve dil bilimin alt dallarından biri olan doğal işleme, insanların kullandıkları dillerin (doğal diller) yapay zekâ yöntemleri kullanılarak işlenmesi, bilgisayarlarla iletişimin sağlanması ve dillerin analiz edilmesini amaçlamaktadır.

Dil bilimi ise, dilleri dilbilgisi, biçimbilimi (morfoloji), söz dizimi (sentaks), ses bilimi (fonetik) vs. açısından inceleyen bilim dalıdır. Doğal dil işleme, dilbilimin ve bilgisayar biliminin beraberliğinden oluşmaktadır. Doğal dil işlemenin birçok uygulaması vardır. Optik karakter okuma, yazım hatası düzeltme, özet çıkarma, bilgi çıkarma, bilgiye erişim, yazılı metinlerden anlam çıkarma, konuşma tanıma, soru-cevap sistemleri, makina çevirisi gibi uygulamaları doğal dil işlemenin çalışma alanı olarak örnek verilebilir. Geçmişten günümüze kadar baktığımızda, doğal dillerin bilgisayarlar tarafından işlenmesinin giderek kolaylaşması ve hızlanmasına rağmen, teknolojinin artması ile elde edilen verilerin artmasından dolayı kullanışlı veriye ulaşmak ve onu işlemek zorlaşmaktadır. Bunun için elde edilen verinin anlamlandırılması veya kullanılabilmesi için diğer bilim alanlarına da ihtiyaç duyulmuştur.

Doğal dil işlemede, (DDİ) tüm veri türlerinin (ses/metin) bilgisayarlar tarafından işlenmesinden önce bir ön işleme adımından geçmesi gerekmektedir. Bu işlemler sayesinde, bozuk olan veya gereksiz olan veriler temizlenir ve veri uygun formata yani önerilen model için kullanılabilir bir hale dönüştürülmüş olur (1). Daha sonra, işlenmeye uygun hale gelen veri bir algoritma ya da model tarafın-

[1] Dr. Öğr. Üyesi Siirt Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, zuleyhayiner@siirt.edu.tr

tır. Yapay sinir ağları yaklaşımların kullanımı diğer bilgisayar uygulamalarında arttığı gibi doğal dil işlemenin de birçok uygulama alanında kullanılmaya başlanmıştır. Bu yaklaşımlar büyük veri kümeleri üzerinde hem zaman hem de başarı noktasında avantaj sağlamışlardır. Bu yaklaşımların, bazı çalışmalarda içerikten bağımsız olarak yüksek başarı elde ettikleri görülmüştür.

KAYNAKLAR

1. R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.
2. E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Computational intelligence magazine,* vol. 9, no. 2, pp. 48-57, 2014.
3. E. Adalı, "Doğal Dil İşleme," *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi,* vol. 5, no. 2, 2012.
4. Y. Kaya and Ö. F. Ertuğrul, "Doküman dili tanıma için yeni bir öznitelik çıkarım yaklaşımı: İkili desenler," *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi,* vol. 31, no. 4, 2016.
5. Y. Kaya, Ö. F. Ertuğrul, and R. Tekin, "Doküman dili tanıma için ikili örüntüler tabanlı yeni bir yaklaşım," *Akademik Bilişim, Eskişehir,* 2015.
6. T. Noyan, F. Kuncan, R. Tekin, and K. Yılmaz, "Döküman dili tanıma için içerik bağımsız yeni bir yaklaşım: Açı Örüntüler," *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi,* vol. 37, no. 3, pp. 1277-1292, 2022.
7. K. Koskenniemi, *Two-level morphology: A general computational model for word-form recognition and production*. University of Helsinki, Department of General Linguistics Helsinki, Finland, 1983.
8. K. Oflazer, "Two-level description of Turkish morphology," *Literary and linguistic computing,* vol. 9, no. 2, pp. 137-148, 1994.
9. A. A. Akın and M. D. Akın, "Zemberek, an open source NLP framework for Turkic languages," *Structure,* vol. 10, no. 2007, pp. 1-5, 2007.
10. "Zemberek Doğal Dil Kütüphanesi." https://github.com/ahmetaa/zemberek-nlp (accessed 13 December 2022.
11. D. Kılınç, A. Özçift, F. Bozyigit, P. Yıldırım, F. Yücalar, and E. Borandag, "TTC-3600: A new benchmark dataset for Turkish text categorization," *Journal of Information Science,* vol. 43, no. 2, pp. 174-185, 2017.
12. R. Dehkharghani, Y. Saygin, B. Yanikoglu, and K. Oflazer, "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis," *Language Resources and Evaluation,* vol. 50, no. 3, pp. 667-685, 2016.
13. M. Eminagaoglu, "A new similarity measure for vector space models in text classification and information retrieval," *Journal of Information Science,* vol. 48, no. 4, pp. 463-476, 2022.
14. A. Üstün, M. Kurfalı, and B. Can, "Characters or morphemes: How to represent words?," 2018: Association for Computational Linguistics.
15. D. Yuret and F. Türe, "Learning morphological disambiguation rules for Turkish," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 2006, pp. 328-334.
16. P. Smit, S. Virpioja, S.-A. Grönroos, and M. Kurimo, "Morfessor 2.0: Toolkit for statistical morphological segmentation," in *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*, 2014: Aalto University.
17. D. Z. Hakkani-Tür, K. Oflazer, and G. Tür, "Statistical morphological disambiguation for agglutinative languages," *Computers and the Humanities,* vol. 36, no. 4, pp. 381-410, 2002.

18. D. Yuret and M. d. l. Maza, "The greedy prepend algorithm for decision list induction," in *International Symposium on Computer and Information Sciences*, 2006: Springer, pp. 37-46.
19. E. Yildiz, C. Tirkaz, H. Sahin, M. Eren, and O. Sonmez, "A morphology-aware network for morphological disambiguation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, vol. 30, no. 1.
20. Q. Shen, D. Clothiaux, E. Tagtow, P. Littell, and C. Dyer, "The role of context in neural morphological disambiguation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 181-191.
21. E. Dayanık, E. Akyürek, and D. Yuret, "Morphnet: A sequence-to-sequence model that combines morphological analysis and disambiguation," *arXiv preprint arXiv:1805.07946,* 2018.
22. Wikipedia. https://en.wikipedia.org/wiki/Hidden_Markov_model (accessed 13 December 2022.
23. B. Merialdo, "Tagging English text with a probabilistic model," *Computational linguistics,* vol. 20, no. 2, pp. 155-171, 1994.
24. C.-H. Chang and C.-D. Chen, "HMM-based part-of-speech tagging for Chinese corpora," in *Very Large Corpora: Academic and Industrial Perspectives*, 1993.
25. N. Saharia, D. Das, U. Sharma, and J. Kalita, "Part of speech tagger for Assamese text," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, pp. 33-36.
26. A. Yajnik, "Part of speech tagging using statistical approach for Nepali text," *International Journal of Cognitive and Language Sciences,* vol. 11, no. 1, pp. 76-79, 2017.
27. D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Third conference on applied natural language processing*, 1992, pp. 133-140.
28. B. Babüroğlu, A. Tekerek, and M. Tekerek, "Türkçe için derin öğrenme tabanlı doğal dil işleme modeli geliştirilmesi," *Web: https://arxiv.org/ftp/arxiv/papers/1905/1905.05699.pdf,* vol. 16, p. 2019, 2019.
29. Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991,* 2015.
30. D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Second ed. Pearson Education, 2009.
31. E. Adalı, "Doğal Dil İşleme," (in scheme="ISO639-1"), *5,* Makaleler(Araştırma) 2016, doi: https://dergipark.org.tr/en/pub/tbbmd/issue/22245/238797.
32. O. Görgün and O. T. Yıldız, "Using morphology in English-Turkish statistical machine translation," in *2012 20th Signal Processing and Communications Applications Conference (SIU)*, 2012: IEEE, pp. 1-4.
33. E. Solak, "On relative clause in Turkish," in *Proceedings of the 7th International Conference on Computer Processing of Turkic Languages (turklang 2019), Simferopol, Russia*, 2019.
34. E. BARUT, "İstatistiksel Makine Çevirisi İle Nöral Makine Çevirisinin Dilbilimsel Parametrelerle Karşılaştırılması: Google Translate," *Akdeniz Havzası ve Afrika Medeniyetleri Dergisi,* vol. 4, no. 1, pp. 103-118, 2022.
35. P. Koehn and R. Knowles, "Six challenges for neural machine translation," *arXiv preprint arXiv:1706.03872,* 2017.
36. S. İlhami, Ü. Hüseyin, and D. HANBAY, "Creating a Parallel Corpora for Turkish-English Academic Translations," *Computer Science,* no. Special, pp. 335-340, 2021.
37. E. Satir and H. Bulut, "Preventing translation quality deterioration caused by beam search decoding in neural machine translation using statistical machine translation," *Information Sciences,* vol. 581, pp. 791-807, 2021.
38. X. Wang, Z. Tu, and M. Zhang, "Incorporating statistical machine translation word knowledge into neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 12, pp. 2255-2266, 2018.
39. C. Amrhein and S. Clematide, "Supervised OCR error detection and correction using statisti-

cal and neural machine translation methods," *Journal for Language Technology and Computational Linguistics (JLCL),* vol. 33, no. 1, pp. 49-76, 2018.

40. J. Moorkens, A. Toral, S. Castilho, and A. Way, "Translators' perceptions of literary post-editing using statistical and neural machine translation," *Translation Spaces,* vol. 7, no. 2, pp. 240-262, 2018.
41. F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research,* vol. 69, pp. 343-418, 2020.