Özyinelemeli sinir ağlarıyla Türkçe varlık ismi tanıma
Recurrent neural networks for Turkish named entity recognition

Yazarlar Gizlenmiştir

***Özetçe* —Bu çalışmada, Türkçe için varlık ismi tanıma görevini çözmek için bir sinir ağı modeli önerilmektedir. Model, cümledeki sözcükleri baştan ve sondan işleyerek cümledeki her bir konumu ayrı ayrı temsil eden birer bağlam vektörü oluşturmaktadır. Bu bağlam vektörü, o konumda bir varlık ismi olup olmadığı hakkında karara varmak amacıyla bir skor vektörü oluşturmak için kullanılır. Sonuçta bu skor vektörleri de koşullu rassal alan (CRF) modelinde kullanılarak son karar verilir. Bu modeli kullanarak yaptığımız deneylerde literatürde bu görevin çözümü için daha önceden önerilen yöntemlere göre daha yüksek bir başarı elde edilmiştir.**

Anahtar Kelimeler—varlık ismi tanıma, sinir ağları, doğal dil işleme.

***Abstract*—In this work, we propose a neural network model for Turkish named entity recognition. Model creates a context vector for every position in the sentence by processing the words in forward and backward directions. This context vector is used to obtain a score vector for deciding whether there is an entity in that position or not. A conditional random field (CRF) model is employed to decide the final entity label. In our experiments using this model, performance results higher than the previous works in the literature were observed.**

Keywords—named entity recognition, neural networks, natural language processing.

I. Giriş

Varlık ismi tanıma (VAT) cümlelerde bulunan varlık atıflarını bulmayı hedefler. Bulunan varlıklar, önceden tanımlı kişi, konum ve kurum gibi sınıflara ayrılırlar. VAT görevi, ilişki tespiti, bilgi tabanı oluşturma ve soru cevaplama gibi görevlerin genellikle ilk aşamalarında kullanılır [1], [2]. Ayrıca arama motorları ve makine çevirisi sistemlerinde de kullanılır [3], [4].

VAT üzerine yapılan ilk çalışmalar elle hazırlanmış kurallar ve kişi, yer ve kurum adlarını içeren özel listelerden yararlanıyorlardı [5], [6]. Bu geleneksel yaklaşımlar genellikle küçük-büyük harf, sözcük uzunluğu, özel sözcük listelerinde bulunup bulunmama ve sözcüğün cümledeki görevi (isim, fiil, zarf, sıfat olup olmaması) gibi elle hazırlanmış özelliklerden faydalanırlar [7], [8]. Varlık ismi tanıma görevini çözmek için geniş bir yelpazeye yerleştirilebilecek birçok makine öğrenimi tabanlı yöntem önerilmiştir. İyi bilinen yaklaşımlardan bazıları koşullu rassal alanlar (CRF) [7], [8], enbüyük entropi [9], saklı anlamsal ilişkilendirme [10] ve karar ağaçları [11] yöntemlerini temel alır.

Literatürdeki son çalışmalara göre, doğal dil cümlelerinin derin öğrenme yöntemleri kullanılarak işlenmesi en yüksek başarıları getirmiştir [12]. Bu başarıyı sağlayan şeylerin başında cümlelerin özyinelemeli sinir ağları [13] kullanılarak işlenmesi ve sözcüklerin sabit uzunluklu vektörler olarak yoğun uzaylarda temsil edilmesi gelir. Bu durum birçok farklı görevde gözlenmiştir: duygu analizi [14], cümle ayrıştırıcı [15], dil modelleme [16], cümle öğesi işaretleme [12]. Bu tür sabit uzunluklu vektörlerin öğrenilmesi ya eğitim sırasında ya da öncesinde Word2Vec [17] veya GloVe [18] gibi yöntemler kullanılarak gerçekleştirilmektedir.

VAT görevini dizi işaretleme sorunu olarak gören çalışmalar bu bulgular üzerine gerçekleştirilmiştir [19]–[22]. Bu çalışmalar genelde VAT görevini çözmek için LSTM (Long short-term memory) veya GRU (gated recurrent unit) modülleri kullanarak cümlenin yapısal ve anlamsal özelliklerini yakalamaya çalışırlar.

Bu çalışmada İngilizce'de en iyi sonuçları almış olan başka bir çalışmanın [19] Türkçe VAT görevinde ne kadar başarılı olduğunu araştırdık. Bunun için Türkçe'de VAT görevi üzerine yapılan çalışmalarda sıklıkla kullanılan bir derlem kullandık [23].

Bu makale şu şekilde bölümlenmiştir: Bölüm II'de deneylerde kullandığımız sinir ağları modeli anlatılmıştır. Bölüm III'de ise bu modelin eğitimiyle ilgili ayrıntıları verdikten sonra eğitim ve test aşamalarında kullandığımız derlem ayrıntılandırılmıştır. Bölüm IV'te sonuçlar ve gelecek çalışmalardan bahsedilmiştir.

978-1-5386-1501-0/18/$31.00

II. MODEL

Verili bir girdi cümlesini $X = (x_1, x_2, \ldots, x_n)$ olarak tanımlıyoruz. Cümlenin i konumundaki her kelime Bölüm II-A'da anlatılan şekilde oluşturulan d uzunluklu bir x_i vektörüyle temsil edilir. Bu vektörler p uzunluklu hücre durumlarına sahip iki LSTM'den [13] oluşan ve Bi-LSTM olarak adlandırdığımız bir modüle ayrı ayrı olarak okunma sırasıyla ve ters sırada iletilir. Böylelikle ileri ve geri yönlerdeki LSTM'lere ait $\overrightarrow{H}$ ve $\overleftarrow{H}$ adını verdiğimiz $n \times p$ büyüklükteki hücre matrisleri elde edilir. Sonuç olarak $\overrightarrow{H}_{i,j}$ ileri LSTM'e ($\overleftarrow{H}_{i,j}$ geri LSTM'e) ait cümledeki i konumunu temsil eden vektörün j boyutunu temsil eder. Her iki yöndeki matris birleştirilerek yeni bir matris elde edilir: $H = [\overrightarrow{H}, \overleftarrow{H}]$. Bu matrisin her i satırı cümledeki i konumu için K nöronlu bir tam bağlı tek aşamalı sinir ağına verilerek ξ_i adı verilen bir varlık tahmin vektörü elde etmek için kullanılır.

Ardarda gelen ve birkaç sözcük boyunca devam eden varlıkları modellemek amacıyla koşullu rassal alan (CRF) yöntemi uygulanır [24]. Literatürde birkaç sözcük boyunca devam eden varlıkları işaretlemek için BIO adı verilen ve varlık etiketlerinin başına B ve I harflerini ekleyen bir notasyon vardır. B harfi varlık etiketinin başını, I harfi ise varlık etiketinin içindeki diğer sözcükleri işaretlemek için kullanılır. Dolayısıyla, bu notasyona göre iki sözcükten oluşan "Ali Kaya" ismine denk gelen etiketler "B-PERSON I-PERSON" olacaktır. Model, B ile başlayan etiketlerin hep ilk sözcükte olmasını hesaba katar. Dahası eğer ikinci sözcük de aynı varlığa aitse I etiketiyle başlamalıdır. Örneğin, I-PERSON etiketinin tahmin olasılığı öncesinde B-PERSON etiketi gelmediği takdirde sıfır olmalıdır. Bu tür bir bilgi sadece ξ_i vektörüne dayanan bir tahmin modeliyle hesaba katılamayacaktır.

Bunu sağlamak için X cümlesi için aşağıdaki hedef fonksiyonu endüşükleştirilmeye çalışılır:

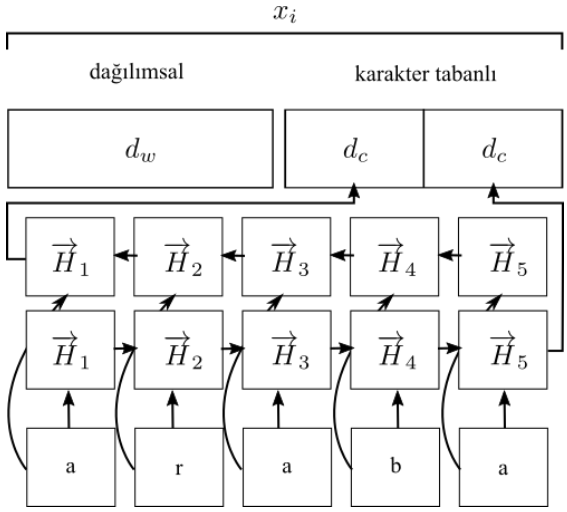
$$s(X, y) = \sum_i A_{y_i, y_{i+1}} + \sum_i \xi_{i, y_i}.$$

Bu cümlede $A_{i,j}$ etiket i'den etiket j'ye geçişi, ξ_i ise i konumundaki varlık tahmin skorlarını temsil eder. Bu modele göre tahmin sırasında en olası varlık etiketleme dizisi $y^* = \arg\max_{\tilde{y}} s(X, \tilde{y})$ denkleminin çözümüyle bulunur.

A. Sözcük vektörleri

Sözcükleri veya cümleleri matematiksel modellere girdi olarak verirken elle hazırlanmış özellikleri kullanmaktansa, sözcükleri sabit uzunluklu vektörler olarak temsil etmenin daha yüksek başarı ölçülerine yol açtığı gösterilmiştir [12], [25]. Bu çalışmada da sözcükler farklı temsil yöntemleriyle elde edilen iki vektörün ardarda eklenmesiyle temsil edildi. Bu yöntemlerin ilki d_w uzunluklu önceden öğrenilmiş sözcük vektörlerini doğrudan almak, ikincisi ise Şekil 1'de gösterilen yöntemle elde edilen $2d_c$ uzunluklu karakter tabanlı vektörler elde etmektir.

Bahsi geçen karakter tabanlı sözcük vektörleri, sözcüklerin yüzey biçimlerindeki karakterlerin sıralamalarını hesaba kattığı için dağılımsal sözcük vektörlerinin yakalayamadığı sözcük içi ilişkileri yakalama şansına sahiptir. Bu, d_c uzunluklu hücrelere sahip iki farklı LSTM'e yüzey biçimindeki karakter sırasıyla ve ters sırada verilerek yapılır. Bu iki LSTM'in son durumlarının ardarda eklenmesi bize o sözcük için karakter tabanlı sözcük vektörünü verir.



Şekil 1: Bi-LSTM kullanarak karakter tabanlı sözcük vektörünün oluşturulması. Ayrıca x_i sözcük vektörünün dağılımsal sözcük vektörüyle de ardarda eklendiği görülüyor.

III. DENEYLER

A. Eğitim

Eğitim sırasında, Bölüm II'deki Bi-LSTM'in parametreleri ile dağılımsal ve karakter tabanlı sözcük vektörlerinin parametreleri öğrenildi. Eğitim öncesi yaptığımız denemelerde, parametre sayısının seçimi hakkında bazı kısa deneyler yaparak her sözcük için dağılımsal sözcük vektörlerinin 100, karakter tabanlı sözcük vektörlerinin 200 uzunlukta olması kararlaştırıldı. Eğitimde kayıp fonksiyonunun türevi her örnek için ayrı ayrı hesaplanarak öğrenme hızı 0.01 seçilerek parametreler güncellendi. Ayrıca çok büyük veya çok küçük türev güncellemelerini durdurmak için güncelleme büyüklüklerini -5 ile +5 sınırı içinde tuttuk. Bunlara ek olarak, sözcük temsillerini Bi-LSTM'e verirken ilgili her parametreyi 0.5 olasılıkla sıfırladık (*dropout* [26].

B. Veriseti

Eğitimleri Türkçe VAT görevi için sıklıkla kullanılan bir derlem [23] ile yaptık. Bu derlemdeki eğitim kümesi 14481 kişi ismi, 9411 konum ismi and 9037 kurum ismi içerirken, test kümesi 1594 kişi ismi, 1091 kurum ismi ve 863 kurum ismi içerir. Eğitimlerde ayrıca 100 uzunluğunda daha önceden öğrenilmiş sözcük vektörleri kullandık. Bunun için atla-gram (*skipgram*) algoritmasını [17] 951 milyon sözcükten oluşan ve 2.045.040 tekil sözcük içeren bir derlem [27] üzerinde çalıştırdık. Bu derlem birçok ulusal gazete, haber sitesi ve kamusal alandaki kitaplardaki Türkçe cümlelerden oluşuyor.

C. Deney sonuçları

Çalışmamız sırasında daha önceden öğrenilmiş sözcük vektörleri kullanmanın en yüksek başarıları verdiğini gördüğümüz için bu bölümdeki sonuçlara sözcük vektörlerinin öncedne öğrenilmesi yerine rassal olarak oluşturulup eğitim sırasında öğrenildiği modelleri dahil etmedik.

Bu çalışma

Sözcük vektörleri		F1-ölçütü
Dağılımsal	Karakter tabanlı	
VAR	YOK	90.96
VAR	VAR	**93.37**

Önceki çalışmalar

Model	F1-ölçüsü
Kuru et al. (2016) [28]	91.30
Demir and Ozgur (2014) [29]	91.85
Seker and Eryigit (2012) [30]	91.94

TABLO I: Yaptığımız deneylerde elde ettiğimiz F-ölçütü sonuçları.

Deney sonuçlarını elde etmek için yaptığımız eğitimlerde her bir eğitim sırasında derlemin en fazla 100 kere üstünden geçtik (*epoch*) ve her geçiş sonunda ayırdığımız bir sağlama kümesindeki cümleler üzerinde F-ölçütünü hesapladık. Bu sağlama kümesinde en iyi sonucu elde eden modelin test kümesi üzerindeki başarısı Tablo I'de görülebilir. Tablonun ilk satırındaki model, sözcükleri temsil ederken sadece dağılımsal önceden öğrenilmiş sözcük vektörleri kullanıyor. Bu modelin %90.96'lık başarısının, tablonun ikinci kısmında belirtilmiş olan önceki çalışmaların başarılarına yakın bir başarı elde ettiği görülüyor. Ancak tablonun ikinci satırında, bu modele karakter tabanlı sözcük vektörlerini de katmanın başarıyı oldukça artırarak önceki çalışmaların başarısının da yukarısına çıkardığı görülüyor. Tablodaki "Önceki çalışmalar" kısmında yer verilen modellerin bizim çalışmamızdan bazı farkları vardır. Örneğin bizim çalışmamızda özel sözcük listeleri kullanılmamışken, Şeker ve Eryiğit ([30]) çalışmasında kullanılmıştır. Tablo I'de verilen başarı buna aittir. Aynı çalışmanın özel sözcük listelerinin kullanılmaması durumundaki başarısı %89.55'e düşmektedir. Öte yandan, Kuru ve ark. ([28]) çalışmasında hiçbir dışsal veri kullanılmadığı not edilmelidir. Demir ve Özgür ([29]) ise elle oluşturulmuş özelliklere dayanan önceden öğrenilmiş dağılımsal sözcük vektörleri de kullanan çok katmanlı bir algılayıcı *perceptron* sistemidir.

IV. Sonuçlar

Bu çalışmada Türkçe varlık ismi tanıma görevi özyinelemeli sinir ağları kullanan bir yöntem ile çözülmüştür. Bu şekilde daha önce literatürde yayımlanmış en yüksek başarılara ulaşılmıştır. İleride yapılacak çalışmalarda, Türkçe gibi biçimbilimsel olarak zengin dillerin sözcüklerinde bulunan kimi özelliklerin başarıma etkisi araştırılacaktır.

Kaynaklar

[1] Y. Liu and F. Ren, "Japanese named entity recognition for question answering system," in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*. IEEE, 2011. [Online]. Available: http://dx.doi.org/10.1109/ccis.2011.6045098

[2] C. Lee, Y.-G. Hwang, and M.-G. Jang, "Fine-grained named entity recognition and relation extraction for question answering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 07*. ACM, 2007. [Online]. Available: http://dx.doi.org/10.1145/1277741.1277915

[3] J. Guo, G. Xu, X. Cheng, and H. Li, "Named entity recognition in query," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2009, pp. 267–274.

[4] B. Babych and A. Hartley, "Improving machine translation quality with automatic named entity recognition," in *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT through Other Language Technology Tools: Resources and Tools for Building MT*. ACL, 2003, pp. 1–8.

[5] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks, "University of Sheffield: Description of the LaSIE-II system as used for MUC-7," in *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. ACL, 1998.

[6] D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, M. Kameyama, D. Martin, K. Myers, and M. Tyson, "SRI International FASTUS system: MUC-6 test results and analysis," in *Proceedings of the 6th Conference on Message Understanding*. ACL, 1995, pp. 237–248.

[7] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, vol. 4. Association for Computational Linguistics, 2003, pp. 188–191.

[8] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proceedings of the 43rd Annual Meeting of ACL*. ACL, 2005, pp. 363–370.

[9] A. E. Borthwick, "A maximum entropy approach to named entity recognition," Ph.D. dissertation, New York University, New York, NY, USA, 1999.

[10] H. Guo, H. Zhu, Z. Guo, X. Zhang, X. Wu, and Z. Su, "Domain adaptation with latent semantic association for named entity recognition," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the NAACL*. ACL, 2009, pp. 281–289.

[11] G. Szarvas, R. Farkas, and A. Kocsor, "A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms," in *International Conference on Discovery Science*. Springer, 2006, pp. 267–278.

[12] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, 1997.

[14] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, Washington, USA: ACL, October 2013, pp. 1631–1642. [Online]. Available: http://www.aclweb.org/anthology/D13-1170

[15] R. Collobert and J. Weston, "A unified architecture for natural language processing," in *Proceedings of the 25th International Conference on Machine Learning - ICML 08*. ACM, 2008. [Online]. Available: http://dx.doi.org/10.1145/1390156.1390177

[16] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[18] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of NAACL-HLT 2016*, 2016, pp. 260–270.

[20] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[21] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," *arXiv preprint arXiv:1603.01354*, 2016.

[22] Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Multi-task cross-lingual sequence tagging from scratch," *CoRR*, vol. abs/1603.06270, 2016.

[23] G. Tür, D. Hakkani-Tür, and K. Oflazer, "A statistical information extraction system for turkish," *Natural Language Engineering*, vol. 9, no. 2, pp. 181–210, 2003.

[24] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning, ICML*, vol. 1, 2001, pp. 282–289.

[25] J. Turian, L.-A. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 384–394. [Online]. Available: http://www.aclweb.org/anthology/P10-1040

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] O. Güngör and E. Yıldız, "Linguistic features in turkish word representations," in *Signal Processing and Communications Applications Conference (SIU), 2017 25th*. IEEE, 2017, pp. 1–4.

[28] O. Kuru, O. A. Can, and D. Yuret, "Charner: Character-level named entity recognition," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, December 2016, pp. 911–921. [Online]. Available: http://aclweb.org/anthology/C16-1087

[29] H. Demir and A. Özgür, "Improving named entity recognition for morphologically rich languages using word embeddings," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, 2014, pp. 117–122.

[30] G. A. Seker and G. Eryiğit, "Initial explorations on using CRFs for Turkish named entity recognition," in *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, December 2012, pp. 2459–2474. [Online]. Available: http://www.aclweb.org/anthology/C12-1150