Bilgisayar Bilimleri ve Mühendisliği Dergisi

Cilt-18
Sayı-1
2025

Gizli Markov Modeli Kullanılarak Özbek Dilinde Sözcüklerin Nitelikleri ile Etiketlemesi

Tagging Words with Their Attributes in Uzbek Language Using the Hidden Markov Model

Botir Boltayevich ELOV
Alisher Navoiy nomidagi Toshkent Davlat O'zbek Tili va Adabiyoti Universiteti
Toshkent, Uzbekistan
elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648

Shahlo Mirdjonovna HAMROYEVA
Alisher Navoiy nomidagi Toshkent Davlat O'zbek Tili va Adabiyoti Universiteti
Toshkent, Uzbekistan
shaxlo.xamrayeva@navoiy-uni.uz
ORCID: 0000-0002-5429-4708

Zilola Yuldashevna XUSAINOVA
Alisher Navoiy nomidagi Toshkent Davlat O'zbek Tili va Adabiyoti Universiteti
Toshkent, Uzbekistan
xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515

Öz

Markov modelleri, doğal dil işlemede en yaygın kullanılan makine öğrenmesi yöntemlerinden biridir. Markov zinciri ve gizli Markov modeli, dinamik sistemleri modellemek için kullanılan stokastik (rastlantısal) yöntemdir ve sistemin mevcut durumunu, önceki durumlara dayanarak tahmin eder. Markov modelleri *konuşma tanıma; konuşma sentezi; biçim bilimsel çözümleyici; makine çevirisi; el yazısını tanıma; sınıflandırma* gibi DDİ uygulamalarında kullanılabilecek karmaşık matematiksel yapılara sahiptir. *Tümcelerin oluşturulmasında doğru bir sözcük dizisi oluşturan Markov zinciri, DDİ çalışmalarında yaygın olarak kullanılmaktadır. Ayrıca bir tümcedeki Varlık Adlarını (NER: VAT) tanımlamak ve gizli Markov modeline dayalı Sözcükleri Nitelikleri ile Etiketlemek (SNE) için kullanılır. Markov modeline dayanarak, gizli etiketler derlemdeki etiketli sözcüklere göre tahmin edilir. Bu makale, Özbek dilinin etiketli derlemini temel alan bir Gizli Markov modeli kullanarak belirli bir tümcenin otomatik SNE etiketlemesine yönelik yöntemler ve algoritmaları sunmaktadır. Özbek dilinde eş anlamlı sözcüklerin bulunduğu ve metnin Sözcükleri Nitelikleri ile etiketlemesinde hangi sözcük kümesine ait olduğunun otomatik olarak belirlenmesi için Viterbi algoritmasının kullanılmasının iyi sonuçlar verdiği görülmüştür. Gizli Morkov modelinde Olasılık değerlerinin yayılımının belirlenmesi adımı önemlidir. Özbek dilinin bu özelliklerine göre metin etiketleme ve Sözcükleri Nitelikleri ile Etiketlemek (SNE) için Gizli Morkov modelinin kullanılması iyi sonuçlar vermektedir. Mevcut çalışmalarla karşılaştırıldığında Gizli Morkov modelinin kullanılması Özbekçe metinlerde Sözcükleri Nitelikleri ile Etiketlemek (SNE) %94'ü doğru çıktı ve Sözcükleri Nitelikleri ile Etiketlemek (SNE) iyileşme yüzdesi yükseldi.*

Anahtar Sözcükler: POS etiketleme, gizli Markov modeli, Markov zinciri, Gizli Markov Modelleri, Stokastik yöntemler, NLP, Kelime sınıfları, Eşsesli sözcük analizi, Geçiş olasılığı, Yayılma olasılığı, Viterbi algoritması.

Abstract

Markov models are one of the most widely used machine learning methods in natural language processing. Markov chain and hidden Markov model are stochastic (random) methods used to model dynamic systems and predict the current state of the system based on previous states. Markov chain, which creates a correct sequence of words in the creation of sentences, is widely used in NLP studies. It is also used to identify NERs in a phrase and for POS tagging based on hidden Markov model. Based on the Markov model, latent labels are predicted based on the labeled words in the language corpus. This article presents methods and algorithms for automatic POS tagging of a given phrase using a hidden Markov model based on a labeled corpus of the Uzbek language.

Keywords: POS tagging, hidden Markov model, Markov chain, Hidden Markov Models, Stochastic methods, NLP, Word classes, Homophone analysis, Transition probability, Propagation probability, Viterbi algorithm.

1.Giriş

Dil biliminin amacı, çevremizdeki sözlü ve yazılı konuşmalarda var olan dilsel özellikleri tanımlamak ve açıklamaktır. Öncelikle insanoğlunun bir diğer görevi de dilin edinimi,

Gönderme, düzeltme ve kabul tarihi: 02.04.2024 - 08.06.2024 – 24.02.2025
Makale türü: İnceleme

anlaşılması ve kullanılmasının bilişsel yönlerini açıklamak olsa da yine konu anlayışımız ile ilgilidir. Dilbilimin önemli görevlerinden biri de dilin dilsel yapısı üzerinden iletişimin nasıl gerçekleştiğini incelemektir. Bu görevi araştırmak için dilsel ifadeleri yöneten bir dizi kural bulunmaktadır.

Daha sonra istatistiksel dil modelleri oluşturulmuş ve DDİ'nin çeşitli alanlarında etkin bir şekilde kullanılmıştır. Uygulamada yararlı olmak, geçerli bir kuram geliştirmekle aynı şey olmasa da istatistiksel dil modellerin etkinliği özgün yaklaşımın geçerliliğini ortaya koymuştur [1].

DDİ çalışmalarında istatistiksel modelleme yöntemlerinden de yararlanılmaktadır. İstatistiksel yöntemler doğal olarak istatistiksel verilerden sonuç çıkarmaya çalışır. Bir olasılık dağılımı tarafından üretilen verileri kullanarak sonuç çıkarmaya çalışır. Makine öğrenmesi açısından Markov modeli, birçok DDİ aracı arasında önemli bir yer tutar. Doğal dil aslında, bağlam içinde anlam elde etmek için sözcük dizilerine dayanan olasılıksal bir yapıdır; dolayısıyla rastantısal modeller arasında Markov modellerinin kullanılmasına uygundur. Bu makale Markov modellerini DDİ çalışmalarına uygulamak için çeşitli çözümler sunmaktadır. Markov modelini anlamak için gerekli olan bazı rastgele, rastantısal ve belirgin süreçler hakkında temel bilgiler aşağıda sunulmuştur [2].

2. Genel bilgiler

2.1. Sözcük Türü Etiketlemesi

Markov Modeli (MM), bilinen olasılıkları analiz ederek gelecekte oluşacak bir şeyin olma olasılığını inceleyen bir yöntemdir. Doğal diller, bağlama uygun anlam elde etmek için sözcüklerin ve deyimlerin sırasına bağlı olan olasılıksal yapılar olarak kabul edilir ve Markov modeli gibi rastantısal modellerin kullanılması olanaklıdır.

Bir tümcedeki her sözcüğün (veya sözcük kümesinin) niteliğine göre etiketlenmesine SNE veya POS adı verilir. İfadeler, biçim bilimsel özellikler veya sözcük etiketleri SNE etiketi olarak işlenebilir. Çoğu durumda adlar, eylemler, önadlar ve benzer sözcük kümeleri SNE etiketi ile kullanılır. Küçük boyutlu dil derlemleri için SNE etiketlemesi insanlar tarafından yapılabilir. Ancak büyük boyutlu dil derleminde SNE etiketleme işlemi çok emek gerektirdiğinden günümüzde bu işlem otomatik yöntemlerle gerçekleştirilmeye çalışılmaktadır. Şu anda, bir otomatik SNE etiketleyici geliştirmek için eğitim derlemi olarak kullanılacak küçük bir derlemde açıklamalar elle girilmeye çalışılmaktadır [3].

SNE etiketleri bir sözcük ve komşuları hakkında önemli bilgiler sağlar. SNE etiketleme veri toplama, ayrıştırma, metni seslendirme (TTS), veri çekme ve derlem çalışmalarında çeşitli amaçlarla kullanılmaktadır. Ayrıca biçim bilimsel çözümleyici, söz dizimsel ayrıştırma, anlamsal ayrıştırma, çeviri ve diğer birçok yüksek düzey DDİ çalışması için bir ara (ilk) adım olarak kullanılmaktadır. Bu durum biçim bilimsel çözümleyici, etiketleme uygulamaları için önemli bir adım haline getirmektedir. Bu makale, Gizli Markov modelini kullanarak Özbekçe metinlerde biçim bilimsel çözümleyici ile etiketleme yöntemlerini tanıtmaktadır. Makale, DDİ'de metin işlemeye istatistik bir yaklaşım olarak Markov modellerine odaklanmaktadır. Çoğu DDİ araştırması, açıklama görevlerine bağımlılığı azaltmak için Markov modellerinin geliştirilmiş kullanımına sahip denetimli modeller kullanmaktadır.

Gizli Markov Modeli (GMM-HMM) gibi güçlü modeller, çok büyük miktarda eğitim verisi gerektirir ve bilinmeyen sözcüğü tanımlama sürecindeki doğruluğu düşüktür. Dünya dillerinin çoğu bu tür modellerin eğitiminde kullanılmak üzere hesaplamayı gerçekleştirmek için yeterli kaynağa (dil derlemleri) sahip değildir. Bu tür dillere yönelik DDİ uygulamaları geliştirme sürecinde pek çok bilinmeyen sözcük ile karşılaşılmaktadır. Bu durum modelin doğruluğunun düşük olmasına neden olmaktadır. Günümüzde doğruluğun arttırılması, küçük kaynaklı diller için önemli bir sorundur. Yazarlar tarafından Aralık 2022 itibarıyla geliştirilen Özbek dilinin etiketli eğitim derlemi 5.000.000.000'dan fazla tümce, yaklaşık 1.200.000 sözcük biçimi içermektedir ve makaledeki örnekler bu derlemden alınan tümcelerdir.

2.2. Gizli Markov Modeli

DDİ görevlerini çözmek için GMM kullanılabilir. GMM'ni incelemek için öncelikle Markov zincirini kısaca anımsamakta yarar vardır [8,9].

Markov zinciri, genellikle durumlar olarak bilinen rastgele değişken dizilerinin olasılıklarını temsil eden bir rastgele süreç modelidir. Her durum belirli bir kümeden değer alabilir. Yani mevcut durumun önceki durumuna bağlı olma olasılığı olarak anlaşılabilir. Bir dizi gözlemlenebilir olayın olasılığını hesaplamak gerektiğinde Markov zinciri kullanılır. Ancak çoğu durumda zincir gizli veya görünmezdir ve her durum, bizim için görülebilen her k gözlemden 1'ini rastgele üretir [10].

Durum, belirli bir zamanda belirli koşullar anlamına gelir. Bize $s_1,s_2,\ldots,s_n$ durumları için bir dizi değişken verilsin. Markov modelinin varsayımına göre:

$$(si = a \mid s1 \ldots si{-}1) = (si = a \mid si{-}1)$$

Markov zincirleri, bir olayın olasılığını, onu farklı bir duruma geçen veya önceki duruma dönen bir durum olarak ele alarak hesaplamak için kullanılır. Markov zincirinin temel açıklaması aşağıdaki Çizelge-1'de verilmiştir:

Çizelge-1. Markov Zincirinin Matematik Gösterimi

$Q = Q_1,Q_2, \ldots Q_n$	N olaydan oluşan bir grup
$A = a_{11},a_{12},\ldots a_{n1}\ldots a_{nn}$	A, her bir $\boldsymbol{aij - i}$ durumundan diğer bir **j** durumuna geçiş olasılığını temsil eden bir olasılık geçiş matrisidir. $\sum_{i=1}^{n} a_{ij} = 1, \quad \forall i$
$\pi = \pi_1, \pi_2,\ldots \pi_n$	S durumları için **başlangıç olasılık dağılımı.** π_i – Markov zincirinin belirli bir i durumunda başlama olasılığıdır. $\sum_{i=1}^{n} \pi_i = 1$

GMM, modellenen sistemin gizli veya gözlemlenmeyen

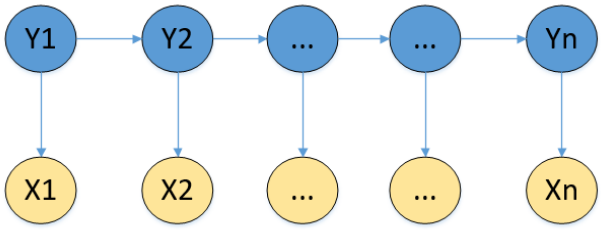
durumlara sahip bir Markov sürecinin analitik bir modelidir. GMM ilk olarak Baum ve Petrie tarafından geliştirilmiş ve ilk olarak DDİ'nin otomatik konuşma tanıma çalışmalarında kullanılmıştır. Jest tanıma, el yazısı tanıma, eşsesli çözümleme, tümce SNE etiketleme gibi makine öğrenimi ve örüntü tanıma gibi DDİ uygulamaları GMM modeline dayanmaktadır. GMM, olasılıksal modelimizde gözlemlenen veya görünen olaylar ve gizli olaylar hakkında akıl yürütmemize olanak tanımaktadır. GMM üç ana adıma dayalı olarak uygulanır: öğrenme, olasılık ve kod çözme:

Öğrenme: Bir dizi eğitim verisi verildiğinde GMM'nin parametrelerinin (özellikle eylem, geçiş ve yayılım olasılıklarının) belirlenmesi;

Olasılık: Eğitilmiş bir GMM'ne dayalı olarak verilen (gözlenen) değişkenler dizisinin olasılığının belirlenmesi;

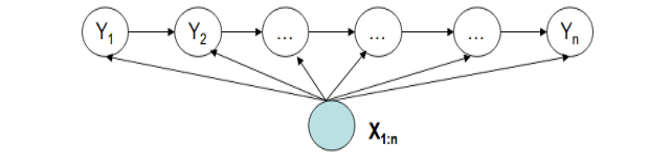
Kod çözme: Eğitilmiş bir GMM ve bir dizi gözlemlenen değişken göz önüne alındığında, gizli durumların olasılıksal bir dizisinin belirlenmesidir.

GMM'nde "*gizli*" sözcüğü, yalnızca sistem tarafından yayılan sinyallerin gözlemlenebileceği anlamına gelir ve kullanıcı durumlar arasındaki rastgele geçişi göremez.



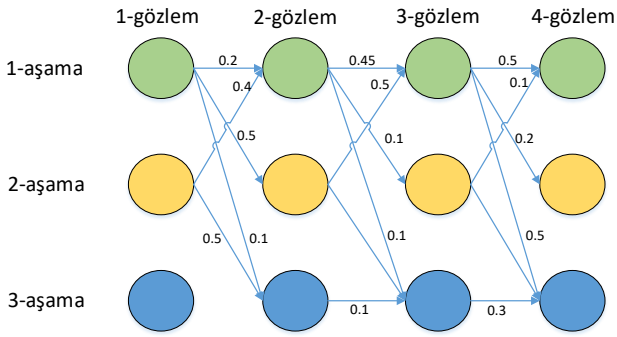
Şekil-1: Gizli Markov modeli

GMM etkili öğrenme algoritmaları ile güçlü bir istatistiksel temele sahiptir ve öğrenme süreci ham dizi verileri üzerinde gerçekleştirilir. Şekil-2'de gösterilen Maksimum Entropi Markov Modelleri (MEMM), komşu durumlar ile gözlemlenen toplam dizi arasındaki ilişkiyi dikkate alır. [14]. Şekil-3'te Markov modelinde gözlemler ve aşamalar gösterilmiştir.



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1},\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}_{1:n}))}{Z(y_{i-1},\mathbf{x}_{1:n})}$$

Şekil-2: Maksimum Entropi Markov modelleri



Şekil-3: Markov modelinde gözlemler ve aşamalar

Markov modelleri, aşağıdaki DDİ uygulamalarında kullanılabilecek matematiksel yapılara sahiptir:

- *konuşma tanıma;*
- *konuşma sentezi;*
- *biçim bilimsel çözümleme;*
- *makine çevirisi;*
- *el yazısını tanıma;*
- *sıra sınıflandırması.*

Kullanıcı bir metni analiz ederken genellikle sözcük kümesine ilişkin etiketleri ayırt etmez, ancak sözcüklerin dizilişinden etiketler hakkında çıkarım yapar. Bu tür etiketler "*gizli*" olarak adlandırılır, çünkü onlar doğrudan izlenemezler [15]. Çizelge-2'de GMM'nin matematik açıklaması verilmiştir.

Çizelge-2. Gizli Markov Modelinin Matematiksel Açıklaması

$S = S_1,S_2,...S_n$	N vakadan oluşan bir grup
$A = a_{11},a_{12},...a_{n1}...a_{nn}$	*A*, her bir ***aij – i*** durumundan diğer bir **j** durumuna geçiş olasılığını temsil eden bir olasılık geçiş matrisidir. $\sum_{i=1}^{n} a_{ij} = 1, \quad \forall i$
$O = O_1,O_2,...O_t$	*T*, tamamı özel bir sözlükten (kaynak) alınmış bir gözlem (O) dizisidir. $V = V_1,V_2,...V_t$
$B = b_i(O_t)$	Tamamı $\mathbf{O_t}$ gözleminin *i* durumundan meydana gelme olasılığını temsil eden bir dizi gözlem olasılığı (yayılma olasılıkları denir).
$P = P_1,P_2,...P_n$	S durumları için **başlangıç olasılık dağılımı.** $\mathbf{P_i}$ – Markov zincirinin belirli bir *i* durumunda başlama olasılığı anlamına gelir. $\sum_{i=1}^{n} p_i = 1$

Bu makale, GMM'ni kullanarak Özbek tümcelerinin SNE etiketlemesine yönelik yöntemler sunmaktadır.

3. Kaynak Taraması

Markov modelleri DDİ'deki çeşitli sorunların çözümünde yaygın olarak kullanılmaktadır. Tek başına kullanılabildiği gibi farklı modellerle birleştirilerek de kullanılabilmektedir. Örneğin, GMM, konuşmaların incelenmesi veya diğer DDİ uygulamalarından dili tanımaya yönelik derin öğrenme ve

sinir ağı modellerini eğitmek için kullanılmaktadır [9].

S. Fawaz, GMM modelinde geçiş ve gözlem matrislerinin oluşturulması, veri dizisinin olasılığının hesaplanması, gizli durumların çıkarılması, GMM'nin profili gibi Markov zincirleri ve ile ilgili bazı yönleri analiz etmiştir.

K. Stratos ve M. Collins tarafından geliştirilen SNE etiketleme algoritması, Brownian kümeleme yöntemi ve Berg-Kirkpatrick logaritmik modeliyle karşılaştırılmaktadır. Araştırmacılar, ayrıca Markov zincirindeki gizli durumları otomatik olarak sözcükler ile eşleştiren bir model geliştirmişlerdir.

A. Kadim ve A. Lazrek, Arapça metinlerin nitelik etiketlemesi (POS) etiketlemesi için Hidden Markov koşut modelini kullanmışlardır. Ana GMM'e ek olarak düşük olasılıklı etiketler için kaynak görevi gören bir yardımcı model de kullanmışlardır. Ayrıca, her iki modeli eğitmek için Nemlar paralel Arapça derlemine dayalı dil bilgisi kurallarından yeni etiketler üretilmiştir. Yaklaşım, Java programlama dili araçları kullanılarak uygulanmış ve bunu değerlendirmek için çeşitli deneyler yapılmıştır.

J. Assunchao, P. Fernandes ve L. Lopez, düzgün tümceleri modellemek için ayrık zamanlı otomatik olarak oluşturulan Markov zincirlerini kullanarak sözcüklerin dil bilgisi sınıflarını tahmin etmek için yöntemler önermişlerdir. Bu yöntemin üstünlüğü hemen hemen her dil, lehçe veya jargon için etkili bir çözüm sunmasıdır.

D. Baishya ve R. Baruah, bir derleme sahip olmayan diller için Biçim Bilimsel Çözümleyici (BBÇ) ile etiketlemenin doğruluğunu arttırmak amacıyla rastantısal bir yöntem ve derin bir öğrenme modeli önermişlerdir. Araştırmaları, dilden bağımsız modellerde bilinmeyen sözcük doğruluğunu ve az miktarda eğitim verisiyle genel doğruluğu iyileştirmeye yönelik yöntemler sunmaktadır. Başlangıçta, eğitim örneklerinin bir parçası olan iki-gramlar ve üç-gramlar, Viterbi algoritması ve GMM kullanılarak bilinmeyen sözcüklerin etiketlenmesinin en yüksek olasılığını hesaplamak için kullanılmıştır. 10K'nın altındaki eğitim veri kümesiyle doğrulukta %12 ila %14'lük bir artış elde edilmiştir. Daha sonra çok az miktarda eğitim verisi ile çalışacak derin sinir ağı modeli geliştirilmiştir. Bu sinir ağı, sözcük düzeyinde, karakter düzeyinde, karakter-iki-gram düzeyinde ve karakter-üç-gram düzeyinde gösterimlere dayalı olup, az miktarda eğitim verisi ile SNE etiketleme işlemi gerçekleştirilmiştir.

Hint dil ailesinden bir dil olan Odia için S. Pattnaik, A. Kumar ve S. Patnaik tarafından bir BBÇ temelli etiketleyici geliştirilmiştir. Etiketleyicinin geliştirilmesi, Viterbi algoritması ile iki-gram GMM'ni temel alıyor ve model, yaklaşık 0,2 milyon lokma büyüklüğünde turizm sektörüne ilişkin bir derlem üzerinde sınanmıştır. 3/1 eğitim-sınama oranıyla önerilen model, sınırlı eğitim boyutundan bağımsız olarak tatmin edici bir başarım göstermiştir.

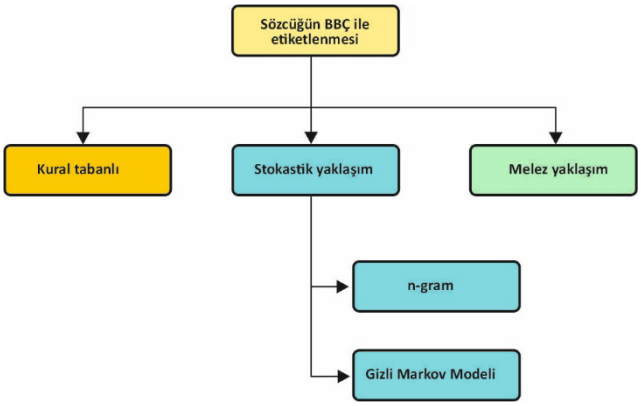
Güney Hint dili (Kannada) için makine öğrenmesi temelli bir SNE etiketleme sistemi S. Atmakuri, B. Shahi ve A. Rao tarafından önerilmiştir. Bu durumda SNE etiketlemeyi gerçekleştirmek için Koşullu Rastgele Alanlar (CRF) seçilir. CRF modeli, biçim bilimsel çözümleyici temelli etiketlemede sıra modelleme, nesne tanıma ve Gizli Markov modellerine karşıt olarak kullanılmıştır. Bu çalışmada eğitim verileri olarak çok büyük üç dil derlemi kullanılmış ve sonuçları karşılaştırılmıştır.

Ch. Alebachev, K. Hivot ve B. Tibebu gibi araştırmacılar, İngilizce Amharca, Afan Oromo, Tigrigna ve diğer diller için SNE etiketleyicileri geliştirip uygulamışlardır. SNE etiketleyicileri geliştirmek için GMM kullanılarak gazeteler (sosyal, ekonomik ve politik yönler dahil), ders kitapları, radyo programları ve bültenler gibi çeşitli kaynaklardan toplanan 1500 tümcelik bir derlem geliştirmişlerdir. Toplanan tümceler, dilbilimciler tarafından her sözcük elle etiketlenmiş ve deneyler aracılığıyla GMM kullanılarak eğitim verileri üzerinde eğitilmiştir. Deneyler, GMM'ne göre SNE etiketlemede %92,77 oranında doğruluk elde edildiğini göstermiştir.

Dim Lam Cing ve Khin Mar Soe tarafından Myanmar dili için Sinir Ağı ve GMM gibi istatistiksel yaklaşımlara dayanan ayrı sözcük bölütleyiciler ve POS etiketleyiciler geliştirilmiştir. Çalışmada Myanmar dilinin karmaşık bir biçim bilimsel yapısı olan sözcük dışında bulunma sorununun üstesinden gelmeye yönelik yöntemler sunulmaktadır.

4. Modern Yaklaşımlar

Konuşmanın etiketlemesi DDİ'de önemli bir araştırma alanıdır. Biçim bilimsel çözümleme işleminde tümcedeki her sözcüğün hangi ulama ilişkin olduğu etiketlenir. Adlar, eylemler, adıllar ve önadlar gibi biçim bilimsel çözümleme etiketleri sözcüklere eklenir ve bilgisayarların tümcelerin anlaşılmasına yardımcı olur. Aşağıdaki Şekil-4'te gösterildiği gibi biçim BBÇ kullanarak etiketleme ile ilgili olarak, uygulamaya yönelik farklı yaklaşımlar ve yöntemler vardır: kural tabanlı, rastantısal veya istatistiksel, melez [24]:



Şekil-4: SNE etiketleme yaklaşımları

Kurallara Dayalı Biçim Bilimsel Çözümleyici ile Etiketleme

Kural tabanlı bir yaklaşım, bir sözcüğü bir etiketle eşleştirmek için bir sözlük kullanır. Bu, veri kümesi açıklamaları veya sözlük oluşturma gibi büyük miktarda el emeği gerektirir. Örneğin, bir önad sözcüğünden sonra sıklıkla bir ad sözcüğü gelir. Bazı durumlarda anahtar sözcüğü tanımlamak için normal ifade şablonları kullanılır. Uzmanlar tarafından derlenen sözlüklerle desteklenen, bilgiye dayalı etiketleyiciler son derece doğru sonuçlar üretir. Ancak kural tabanlı SNE etiketlemenin uygulanmasında birtakım sınırlamalar vardır [24].

Rastantısal Yöntemleri Kullanarak BBÇ ile Etiketleme

Bu yöntemle etiketleme işlemi otomatik olarak gerçekleştirilir ve veri kodlama ve sözlük oluşturma gerektirmez. Bir sözcüğü etiketlemek için istatistik, sıklık ve olasılık ilişkilerinden yararlanılır. Etiketleme işleminde bir sözcüğün belirli bir etiketle ilişkilendirilme olasılığı veya ardışık sözcüklerin sıklığı belirlenir. Bu yaklaşımda olasılıklar, eğitim derlemindeki etiketli verilere dayanarak hesaplanır. Rastantısal olarak SNE etiketleme, n-gramlar veya gizli Markov modelleri kullanılarak gerçekleştirilir [25].

Üç yaygın n-gram türü vardır: uni-gram (tek sözcük), bi-gram (iki sözcük) ve tri-gram (üç sözcük). Aşağıdaki Çizelge-3'te "Bir kurgu kitabı okudum" tümcesine karşılık gelen n-gramlar listelenmiştir:

Çizelge-3: Tümceye uygun gelen n-gramlar

N-gramlar	Sözcük/sözcük örnekleri
Uni-gram	Men badiiy kitobni o'qidim
Bi-gram	Men badiiy badiiy kitobni kitobni o'qidim
Tri-gram	Men badiiy kitobni badiiy kitobni o'qidim

Bu yaklaşım, n-gram olasılığını hesaplamak için istatistiksel bir model kullanır ve belirtilen n-gramlara karşılık gelen bir etiket atar. Bir uni-gram sözcüğün olasılığı aşağıdaki denklemle belirlenir:

$$P(g_i \mid d_i) = \frac{freq(d_i|g_i)}{freq\,(d_i)} \quad (1)$$

Bigram sözünün olasılığı aşağıdaki denklemle belirlenir:

$$P(g_i \mid d_i) = P(d_i \mid g_i) * P(g_i \mid g_{i-1}) \quad (2)$$

Trigram sözünün olasılığı aşağıdaki denklemle belirlenir:

$$P(g_i \mid d_i) = P(d_i \mid g_i) * P(g_i \mid g_{i-2}, g_{i-1}) \quad (3)$$

Burada g bir etiket ve w bir sözcük dizisidir. Geçerli sözcüğün geçerli etikete göre olasılığını ve geçerli etiketin önceki etikete göre olasılığını ifade eder. BBÇ ile etiketlemede GMM, metindeki her sözcüğe karşılık gelen bir etiketle ilişkilendirir. POS etiketleri gizli durumlar olarak kabul edilir ve GMM derlemde gözlemlenen (etiketli) sözcüklere bağlı olarak etiketi tahmin etmeye çalışır. Aşağıdaki eşitliği sağlayacak şekilde tanımlamak gerekir:

$$\prod_{i=1}^{n} (d_i \mid g_i) * P(g_i \mid g_{i-1}) \quad (4)$$

Melez Yöntemlere Dayalı BBÇ ile Etiketleme

Melez bir yöntem, BBÇ ile etiketlemede kural tabanlı ve rastantısal yaklaşımları birleştirir. İlk adımda model istatistiksel yöntemler kullanılarak eğitilir. Daha sonra sonucun verimliliğini artırmak için kural tabanlı bir yaklaşım uygulanır [26].

3.1 POS (Sözcük Türü Etiketlemesi) Etiketleme Yöntemleri

Özbekçe metinlerin POS etiketlemesi için kullanılabilecek farklı yöntemler vardır:

Kurallara Dayalı BBÇ ile Etiketleme

Kural tabanlı POS etiketleme modelleri, bir dizi dil bilgisi kuralını kullanarak metindeki kelimelere POS etiketlerini atar. Bu dil bilgisi kurallarına genellikle bağlam çerçevesi kuralları denir. İşte bu tür kurallara örnekler:

"Belirsiz/bilinmeyen sözcük "-di" ekiyle bitiyorsa, eylem olarak işaretleyin."

Dönüşüm Tabanlı Etiketleme

Dönüşüm tabanlı yaklaşımlar, önceden insan tarafından tanımlanmış kuralların yanı sıra eğitim sırasında geliştirilen otomatik olarak uygulanan kuralları kullanır.

Derin Öğrenme Modelleri

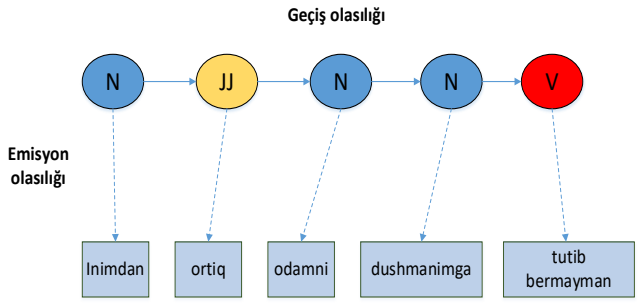
POS etiketleri için farklı derin öğrenme modelleri kullanılmaktadır. Örneğin Meta-BiLSTM modeli yaklaşık %97 doğruluk sağlamaktadır [27,28].

Rastantısal (olasılıksal) Etiketleme

Olasılıksal bir yaklaşım sıklık, olasılık veya istatistikleri içerir. En basit olasılıksal yaklaşım, eğitim verilerinde belirli bir sözcük için en sık kullanılan etiketi belirler ve bu bilgiyi o sözcüğü düz metin olarak etiketlemek için kullanır. Ancak bazı durumlarda bu yaklaşım dilin dil bilgisi kurallarına uymayan etiketler üretebilir. Böyle bir yaklaşım, bir tümce için farklı olası etiket dizilerinin olasılıklarını hesaplamak ve POS etiketlerini en yüksek olasılıkla diziden atamaktır. GMM POS etiketleme için olasılık temelli bir yaklaşımdır.

Gizli Markov Modeli ile SNE Etiketleme

GMM, rastantısal yöntemler kullanarak metinde BBÇ ile etiketlemeye yönelik bir modeldir. GMM konuşma, yazma, jest tanıma ve biyo-informatik gibi DDİ uygulamalarına uygulanabilir. Luis Serrano tarafından önerilen bir örneği ele alalım ve Gizli Markov Modeli kullanarak metin için uygun bir etiket dizisi seçelim [29,30]. Durum Şekil-5'te gösterilmiştir.



Şekil-5: Luis Serrano tarafından önerilen HMM yöntemi

Bu örnekte yalnızca adlar, kipler ve eylemlerden oluşan 3 SNE etiketi görülür. "*Doğa gelin gibi güzeldir*" tümcesi ad, nesne, ad ve önad olarak etiketlenerek bu dizinin geçiş olasılığını ve yayılma olasılığını hesaplamak gerekir.

Geçiş olasılığı belirli bir dizinin olasılığıdır; örneğin bir adın ardından bir ilgeç gelmesi, bir önadın ardından bir ilgeç

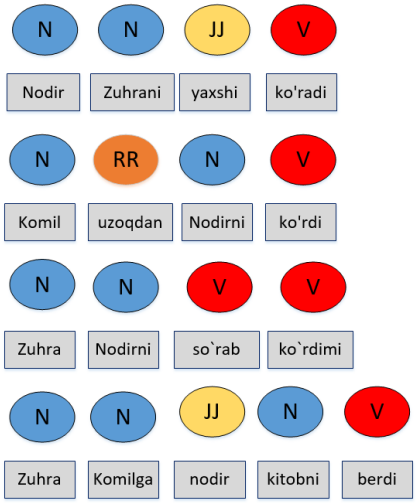
gelmesi ve bir önadın ardından bir adın gelmesi olasılığı. Bu tür olasılığa geçiş olasılığı denir. Belirli bir dizinin doğru olması için bu olasılık değerinin yüksek olması gerekir.

Şimdi "*tabiat*" sözcüğünün ad, "*gibi*" sözcüğünün ad, "*gelin*" sözcüğünün ad, "*serbezak*" sözcüğünün de önad olma olasılığını hesaplayalım. Bu olasılıklar dizisine emisyon (yayılma) olasılığı denir.

Yukarıdaki iki olasılığı aşağıdaki tümce kümesi için hesaplanmıştır:

1. **Nadir Zuhra'yı** seviyor.
2. **Kamil** uzaktan **Nadir'i** gördü.
3. **Zuhra Nadir'e** sordu mu?
4. **Zuhra** nadir kitabı Kamil'e verdi.

Nadir, **Zuhra** ve **Kamil**'in insan özel adları olduğuna dikkat edelim.



Şekil-6: SNE etiketlenen sözcükler

Yukarıdaki tümcelerde *Nadir* sözcüğü ad olarak üç kez ve önad olarak bir kez geçmektedir. Yayılım olasılığını hesaplamak için benzer şekilde bir hesaplama çizelgesi, aşağıda Çizelge-4'te oluşturulmuştur:

Çizelge-4: Verilen Tümcelere Karşılık Gelen Etiket Sıklıkları

Sözcükler	Ad	Önad	İlgeç	Eylem
Nodir	3	1	0	0
Zuhra	3	0	0	0
Komil	2	0	0	0
yaxshi	0	1	0	0
ko'radi/ko'rdi/ ko'rdimi	0	0	0	3
uzoqdan	0	0	1	0
so'rab	0	0	0	1
kitobni	1	0	0	0
berdi	0	0	0	1

Çizelge-4'teki değerler verilen dört tümceye ilişkin geçiş olasılık değerleridir. GMM, yukarıdaki tablolardan belirli bir tümce için uygun etiket sırasını nasıl belirlediği şöyle açıklanır: Yeni bir tümce alınıp Çizelge-5'te gösterildiği gibi yanlış etiketlerle işaretlenir.

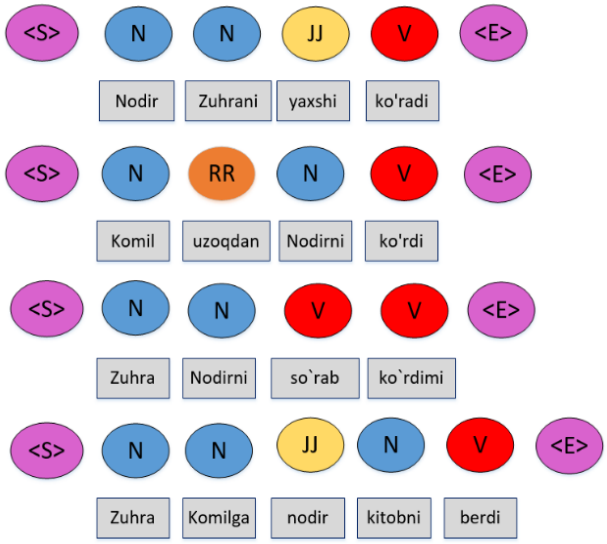
Çizelge-5: Yayılma Olasılığı Değerleri

Sözcük	Ad (N)	Önad (JJ)	İlgeç (RR)	Eylem (V)
Nodir	3/9	1/2	0	0
Zuhra	3/9	0	0	0
Komil	2/9	0	0	0
Yaxshi	0	1/2	0	0
ko'radi/ko'rdi/ ko'rdimi	0	0	0	3/5
Uzoqdan	0	0	1	0
so'radimi	0	0	0	1/5
Kitobni	1/9	0	0	0
Berdi	0	0	0	1/5

Çizelge- 4 ve Çizelge-5'deki değerlere dayanarak aşağıdaki sonuçlara varılır:

- ***Zuhra** sözcüğünün ad olma olasılığı = 3/9*
- ***Zuhra** sözcüğünün belirteç olma olasılığı = 0*
- ***Nodir** sözcüğünün ad olma olasılığı = 3/9*
- ***Nodir** sözcüğünün önad olma olasılığı = 1/2*

Kalan tüm olasılıklar da yukarıdaki yöntemle belirlenebilir. Bu emisyon olasılığıdır. Bir sonraki adımda geçiş olasılığını hesaplamak gerekir. Bu nedenle iki ek <S> ve <E> etiketi tanımladık. Aşağıdaki Şekil-7'te gösterildiği gibi, her cümlenin başına <S> ve sonuna <E> yerleştirilir:



Şekil-7: Sözcüğün biçim bilimsel sınıfına göre etiketlenmiş tümceler

Bir sonraki adımda etiket sayısını temsil eden Çizelge-6 oluşturulur:

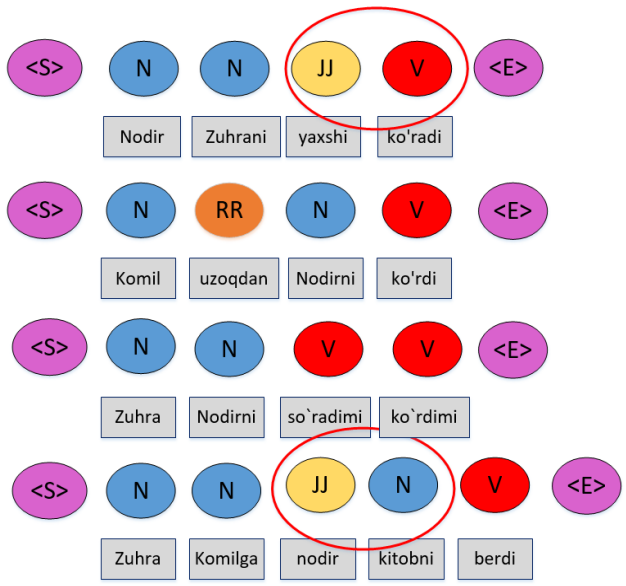
Çizelge-6: Sözcük Kümelerinin Ardışık Oluşum Olasılıkları

	N	JJ	RR	V	<E>
<S>	4	0	0	0	0
N	3	2	1	2	0
JJ	1	0	0	1	0
RR	1	0	0	0	0
V	0	0	0	1	4

Yukarıdaki Çizelge-6'da <S> etiketinin ardından N etiketinin dört kez geldiğini görüyoruz. Önad etiketi (JJ) İsim etiketinden

(N) sonra yalnızca 1 kez geçer. Bu nedenle ikinci yazı 1'e eşittir. Çizelgenin geri kalanı da aynı şekilde doldurulur.

Bir sonraki adımda çizelgenin satırındaki her terim dikkate alınan etiketin toplam sayısına bölüyoruz. Örneğin önad etiketinin (JJ) ardından 2 kez başka bir etiket gelir. Bu nedenle her bir öğe Şekil-8'de gösterildiği gibi bölünür:



Şekil-8: Sözcüğün biçim bilimsel kümesine göre etiketlenmiş tümceler.

Çizelge-7'de yayılma olasılık değerleri gösterilmiştir.

Çizelge-7: Geçiş Olasılık Değerleri

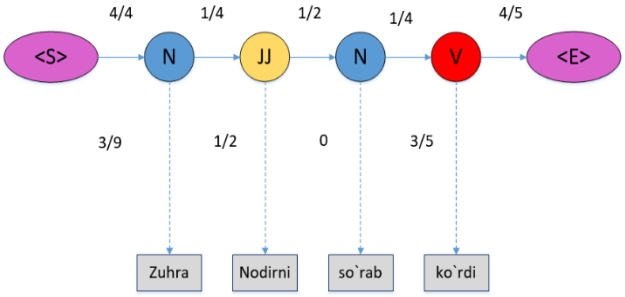
	N	JJ	RR	V	<E>
<S>	4/4	0	0	0	0
N	3/8	2/8	1/8	2/8	0
JJ	1/2	0	0	1/2	0
RR	1/2	0	0	0	0
V	0	0	0	1/5	4/5

Çizelge-7'deki değerler, verilen dört tümceye ilişkin geçiş olasılık değerleridir. GMM, Yukarıdaki çizelgelerden seçilen bir tümce için uygun etiket sırasının nasıl oluşturulacağını belirlemeye çalışır. Bu süreç şöyle açıklanır: İlk aşamada tümce yanlış olarak etiketlenir:

"*Zuhra Nadir'e sordu*" tümcesi (yanlış olarak) şöyle etiketlenir:

- *Zuhra – ad;*
- *Nodirni – önad;*
- *so'rab – ad;*
- *ko'rdi – eylem.*

Bir sonraki adımda bu dizinin doğru olma olasılığını şu şekilde hesaplıyoruz:

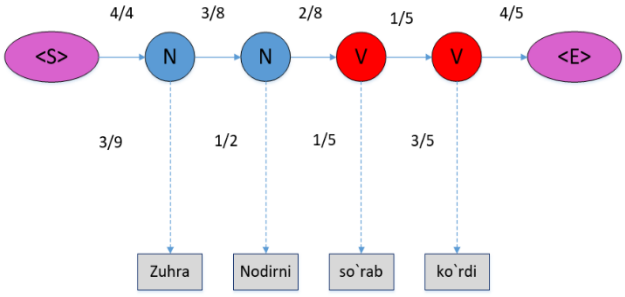


Şekil-9: Sözcük kümelerinin sırayla gelme olasılığı

Çizelge-7'de görüldüğü gibi önad etiketinin gelme olasılığı, ad (N) etiketinden sonra 1/4'tür. Ayrıca, *Zuhra* sözcüğünün Ad olma olasılığı da 3/9'dur. Aynı şekilde grafikteki her olasılık hesaplanır. Bu olasılıkların çarpımı dizinin doğru olma olasılığını belirler. Etiketler doğru olmadığından (biçimlendirilmediğinden) çarpan sıfırdır:

$$\frac{4}{4}\cdot\frac{3}{9}\cdot\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{1}{1}\cdot 0\cdot\frac{1}{4}\cdot\frac{3}{5}\cdot\frac{4}{5}=0$$

Verilen tümcedeki sözcükler doğru şekilde etiketlenmişse, aşağıda gösterildiği gibi sıfırdan büyük bir olasılık üretilir:

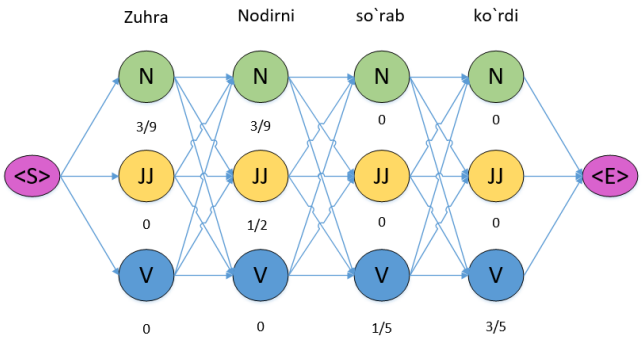


Şekil-10: Sözcük kümelerinin ardışık olarak ortaya çıkma olasılığı

Verilen tümcedeki sözcük sırasına karşılık gelen geçiş ve yayılma olasılık değerlerinin çarpımını hesaplanır:

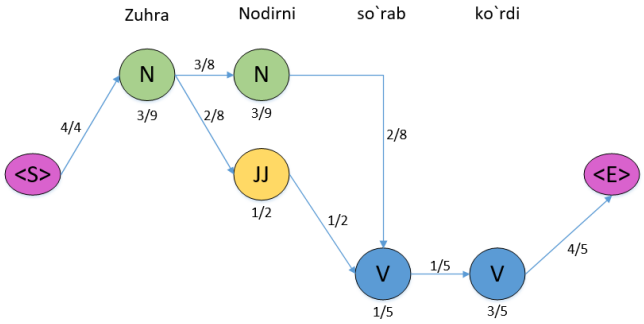
$$\frac{4}{4}\cdot\frac{3}{9}\cdot\frac{3}{8}\cdot\frac{1}{2}\cdot\frac{2}{8}\cdot\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}\cdot\frac{4}{5}=0.0003$$

Örneğin, bahsettiğimiz üç SNE etiketi göz önüne alındığında 16 farklı etiket kombinasyonu oluşturulabilir. Bu durumda 16 kombinasyonun tümünün olasılıklarını hesaplamak olanaklı görünmektedir. Ancak hedef daha büyük tümceleri tanımlamak olduğunda ve Penn Treebank projesindeki tüm SNE etiketleri dikkate alındığında olası kombinasyonların sayısı katlanarak artacak ve hedef olanaksız olacaktır. Bu 16 kombinasyonu yol olarak düşünüp ve grafiğin köşe ve kenarlarının her birine Şekil-11'de gösterildiği gibi geçiş ve yayılma olasılıklarını işaretleyelim.



Şekil-11: Belirli bir tümceyle eşleşen tüm SNE etiketlerinden oluşan (genel) bir grafik

Bir sonraki adımda grafiğin sıfır olasılığa sahip tüm köşe ve kenarları grafikten çıkarılmalıdır. Ayrıca, uç noktaya gitmeyen köşeler de kaldırılır. Böylece Şekil-12'deki sonuç elde edilir.



Şekil-12: Belirli bir tümceye karşılık gelen SNE etiketlerinin grafiği

Başlangıç noktasından bitiş noktasına kadar yalnızca iki yol vardır ve her bir yola ilişkin olasılık şöyle hesaplanır:

$$< \mathbf{S} >\rightarrow \mathbf{N} \rightarrow \mathbf{N} \rightarrow \mathbf{V} \rightarrow \mathbf{V} \rightarrow < \mathbf{E} > = \frac{4}{4}\cdot\frac{3}{9}\cdot\frac{3}{8}\cdot\frac{3}{9}\cdot\frac{2}{8}\cdot\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}\cdot\frac{4}{5} = 0.0002$$

$$< \mathbf{S} >\rightarrow \mathbf{N} \rightarrow \mathbf{JJ} \rightarrow \mathbf{V} \rightarrow \mathbf{V} \rightarrow < \mathbf{E} > = \frac{4}{4}\cdot\frac{3}{9}\cdot\frac{2}{8}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}\cdot\frac{4}{5} = 0.0004$$

Yukarıdaki hesaplamalarda ikinci sıranın gelme olasılığı çok daha yüksektir ve dolayısıyla GMM tümcelerdeki sözcüğü bu sıraya göre etiketler. Ama sonuç beklediğimiz gibi çıkmamıştır, yani "*Zühra Nadirni'yi gördü*" tümcesindeki "*Nadirni*" sözcüğü önad olarak etiketlenmiştir. Markov modelinin Viterbi algoritması kullanılarak optimize edilmesi istenilen sonucun elde edilmesini sağlayabilir. GMM algoritmasının sözde kodu aşağıda verilmiştir:

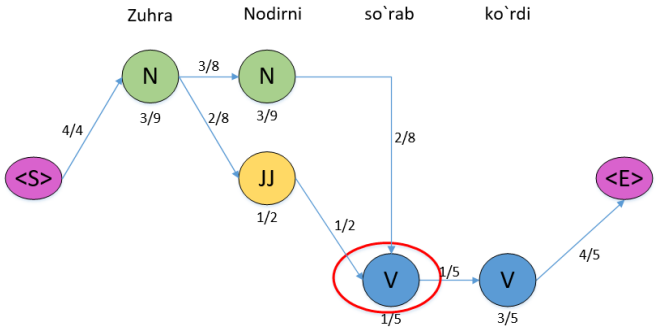
Yashirin Markov modeli algoritması

Initialize:
    $\sigma \leftarrow k \times N$ *array*
For s = 1 … k: $\sigma[1, s] \leftarrow \pi(s) Pr[O^1 | s]$
    $X \leftarrow k \times N$ *array*
Populate σ and X
    For i = 2 … *N*:
        For s = 1 … *k*:
            $xprev \leftarrow \text{argmax}_x \{\sigma[i\text{-}1, x] \Pr[x \rightarrow s]\}$
            $X[i, s] \leftarrow xprev$
            $\sigma[i, s] \leftarrow \sigma[i\text{-}1, xprev] \Pr[x \rightarrow s] Pr[O^i | s]$
Reconstruct OptPath:
    $s \leftarrow \text{argmax}_x \{\sigma[N, x]\}$
Optpath ← *EmptyList*
    For j = *N* … 1:
        Optpath ← s :: *Optpath*
        if j > 1: $s \leftarrow X[j, s]$
Return OptPath

3.2 Viterbi Algoritmasını Kullanarak GMM Optimizasyonu (Eniyilemesi)

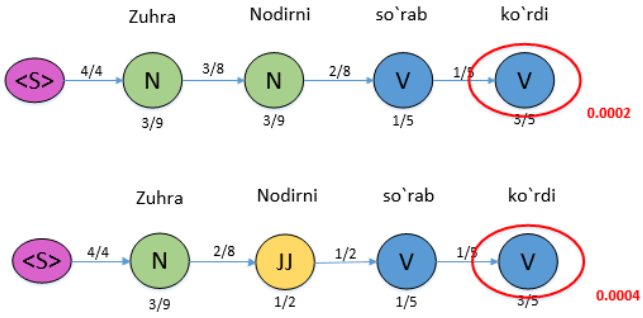
Viterbi algoritması, Viterbi yolu olarak adlandırılan ve gözlemlenen olay dizilerini, özellikle Markov veri kaynaklarını ve GMM kullanan bir dizi gizli durumu tanımlamak için dinamik bir programlama algoritmasıdır.

Yukarıdaki yorumlarda GMM en iyi duruma getirilmiş ve hesaplamalar 16'dan sadece ikiye indirilmiştir. Bir sonraki adımda, Viterbi algoritmasını kullanarak Gizli Markov modelinin daha da optimizasyonuna yönelik yöntemler sunulmaktadır. Daha önce kullandığımız örneği kullanarak Viterbi algoritması buna uygulanmıştır. Şekil-13.



Şekil-13: Viterbi algoritmasıyla uyumlu SNE etiketlerinin grafiği

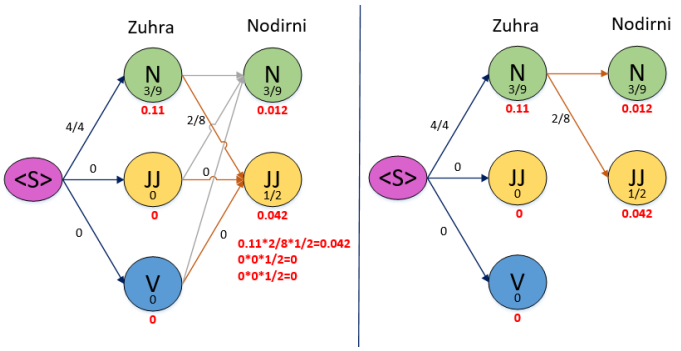
Şekil-13'teki örnekte, bölünmüş bir grafiğin tepe noktasını ele alınmaktadır. Şekil-14'te gösterildiği gibi bu zirveye giden iki yol vardır:



Şekil-14: Belirli bir tümceye karşılık gelen SNE etiketlerinden oluşan yollar

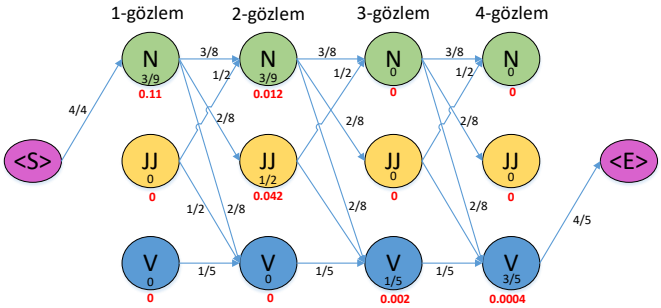
Bir sonraki adımda en düşük olasılığa sahip yol analiz edilir. Bu durum Şekil-15'te gösterildiği gibi grafikteki tüm durumlar için

aynı yöntem izlenir:



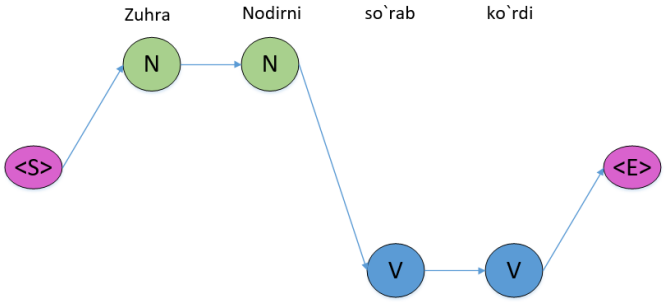
Şekil-15: SNE etiketlerinin grafik yollarının olasılık değerleri

Bir sonraki adımda grafiğin kenarına giden tüm yolların olasılıkları hesaplandıktan sonra olasılık değeri daha düşük olan kenarların veya yolun çıkarılması gerekir. Şekil-15'ten bazı köşelerin sıfır olasılığa sahip olduğu da görülebilmektedir. Grafiğin sonuna giden tüm yolların olasılıklarını hesapladıktan sonra Şekil-16'daki grafik oluşturulur.



Şekil-16: SNE etiketleri grafiğindeki yolların olasılık değerleri (genel)

En uygun yolu elde etmek için son üçte birlik kısımdan geriye doğru çalışırız, bu da bize aşağıda gösterilen yolu verir:



Şekil-17: Viterbi algoritmasına göre tanımlanan SNE etiketlerinin grafiği

Bu algoritma, iki yol sunan önceki yönteme kıyasla yalnızca bir yol döndürür. Dolayısıyla bu algoritma kullanılarak daha az hesaplama yapılır. Viterbi algoritması uygulandıktan sonra model tümce şu şekilde etiketlenir:

- *Zuhra – ad;*
- *Nodirni – ad;*
- *so'rab – eylem;*
- *ko'rdi – eylem.*

Bunlar geçerli etiketlerdir ve GMM'nin sözcüklere karşılık gelen SNE etiketleriyle başarılı bir şekilde etiketleyebildiği sonucuna varılabilir.

Sonuç ve Öneriler

GMM, bir olayın olasılığını analiz etmek için tasarlanmış grafiksel bir modeldir. Bu modelde kullanılan algoritmalar rastgele süreçleri incelemek ve çıkarmak için kullanılır. Bu çalışma aşağıdaki çalışmaları kapsamaktadır.

1. GMM ile Özbek dilinin etiketli ulusal derlemindeki gözlem verilerine dayanarak, verilen tümcelerin gizli hallere göre koşullu dağılımı belirlenebilir ve en yüksek olasılığa sahip değere göre sözcük etiketlemesi yapıldı.
2. GMM algoritmasının etkinliğini arttırmak için Viterbi algoritması kullanılarak gizli durumların sırasının Viterbi yolu şeklinde belirlendi.
3. Özbek dili derlemindeki 4 etiketli tümce örneği üzerinden, verilen tümcelerin otomatik etiketlenme işlemi birbirini izleyen adımlar esas alınarak gerçekleştirilmiş ve analiz sonuçları grafikler aracılığıyla sunulmuştur.
4. SNE etiketlemenin kalitesini artırmak için gözlem verilerinin (derlemdeki etiketli cümleler) miktarının artırılmasının gerektiği gösterilmiştir.
5. Özbek dili metinleri analizin kalitesi ve verimliliğini daha artırmak için Baum-Welch algoritması kullanıldı.

Kaynakça

[1] Rohman, Y. A., Kusumaningrum, R., *Twitter Storytelling Generator Using Latent Dirichlet Allocation And Hidden Markov Model POS-TAG (Part-Of-Speech Tagging)*. ICICOS 2019 - 3rd International Conference on Informatics and Computational Sciences: Accelerating Informatics and Computational Research for Smarter Society in The Era of Industry 4.0, Proceedings. https://doi.org/10.1109/ICICoS48119.2019.8982411, (2019).

[2] Muljono, U., Supriyanto, C., *Morphology Analysis For Hidden Markov Model Based Indonesian Part-of-Speech Tagger*. Proceedings - 2017 1st International Conference on Informatics and Computational Sciences, ICICoS 2017, 2018-January. https://doi.org/10.1109/ICICOS.2017.8276368, (2017).

[3] Xusainova, Z. Y. , *NLP: Tokenizatsiya, Stemming, Lemmatizatsiya Va Nutq Qismlarini Teglash. O'Zbek Amaliy Filologiyasi İstiqbollari / Respublika İlmiy-Amaliy Konferensiya To'Plami*. Elektron nashr / ebook. – Toshkent: ToshDO'TAU, 26.10.2022. 159-163 b.

[4] Baishya, D., Baruah, R., *Resource Languages with Improved Hidden Markov Model and Deep Learning*. International Journal of Advanced Computer Science and Applications, 12(10). https://doi.org/10.14569/IJACSA.2021.0121011, (2021).

[5] Pattnaik, S., Nayak, A. K., Patnaik, S., *A Semi-Supervised Learning Of HMM To Build A POS Tagger for a Low Resourced Language*. Journal of Information and Communication Convergence Engineering, 18(4). https://doi.org/10.6109/jicce.2020.18.4.207, (2020).

[6] Elov, B., Hamroyeva, Sh., Axmedova, X., *Methods for Creating a Morphological Analyzer*, 14th International Conference on Intellegent Human Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent. (2022).

[7] Assunção, J., Fernandes, P. Lopes, L., *Language Independent Pos-Tagging Using Automatically Generated Markov Chains*. Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE, 2019-July. https://doi.org/10.18293/SEKE2019-097, (2019).

[8] Talarico, E., Leão, W., Grana, D., *Comparison of Recursive Neural Network and Markov Chain Models İn Facies Inversion*. Mathematical Geosciences, 53(3). https://doi.org/10.1007/s11004-020-09914-w, (2021).

[9] Myers, D. S., Wallin, L., Wikström, P., *An Introduction to Markov Chains and Their Applications Within Finance.* Mathematical Sciences-Chalmers University of Technology and University, (2017).

[10] Cahyani, D. E., Mustikaningtyas, W., *Indonesian Part of Speech Tagging Using Maximum Entropy Markov Model on Indonesian Manually Tagged Corpus.* IAES International Journal of Artificial Intelligence, 11(1). https://doi.org/10.11591/ijai.v11.i1.pp336-344 , (2022).

[11] Jurafsky, D., Martin, J., *Speech and Language Processing*. In Speech and Language Processing. (Vol. 3). (2014).

[12] Cing, D. L., Soe, K. M. , *Improving Accuracy of Parto-of-Speech (POS) Tagging Using Hidden Markov Model and Morphological Analysis For Myanmar Language.* International Journal of Electrical and Computer Engineering, 10(2). https://doi.org/10.11591/ijece.v10i2.pp2023-2030 , (2020).

[13] Kadim, A. Lazrek, A., *Parallel HMM-Based Approach for Arabic Part of Speech Tagging.* International Arab Journal of Information Technology, 15(2). (2018).

[14] Suleiman, D., Awajan, A., Etaiwi, W., *The Use of Hidden Markov Model in Natural ARABIC Language Processing: A Survey.* Procedia Computer Science, 113. https://doi.org/10.1016/j.procs.2017.08.363 , (2017).

[15] Kumar, A.. Katiyar, V., Kumar, P., *A Comparative Analysis of Pre-Processing Time in Summary Of Hindi Language Using Stanza and Spacy.* IOP Conference Series: Materials Science and Engineering, 1110(1). https://doi.org/10.1088/1757-899x/1110/1/012019, (2021).

[16] Atmakuri, S., Shahi, B., Ashwath Rao, B., Muralikrishna, S. N. , *A Comparison of Features For POS Tagging in Kannada*. International Journal of Engineering and Technology(UAE), 7(4). https://doi.org/10.14419/ijet.v7i4.14900 (2018).

[17] Chiche, A., Kadi, H., Bekele, T., *A Hidden Markov Model-Based Part of Speech Tagger for Shekki'noono Language.* International Journal of Computing, 2021(4). https://doi.org/10.47839/ijc.20.4.2448, (2021).

[18] Al-Anzi, S., A.Zeina, D., *Statistical Markovian Data Modeling for Natural Language Processing*. International Journal of Data Mining & Knowledge Management Process, 7(1). https://doi.org/10.5121/ijdkp.2017.7103 , (2017).

[19] Kumawat, D., Jain, V., *POS Tagging Approaches: A Comparison.* International Journal of Computer Applications, 118(6). https://doi.org/10.5120/20752-3148, (2015).

[20] Jassim, A. K., Al-Bayaty, B. , *A Stochastic Approach to Identify POS in Iraqi National Song Using N-Iterative HMM Using Agile Approach.* IOP Conference Series: Materials Science and Engineering, 1094(1). https://doi.org/10.1088/1757-899x/1094/1/012019, (2021).

[21] Hadni, M., Alaoui Ouatik, S., Lachkar, A., Meknassi, M., *Hybrid Part-of-Speech Tagger For Non-Vocalized Arabic Text.* International Journal on Natural Language Computing, 2(6). https://doi.org/10.5121/ijnlc.2013.2601, (2013).

[22] Bohnet, B., McDonald, R., Simões, G., Andor, D., Pitler, E., Maynez, J., *Morphosyntactic Tagging with a Meta-Bilstm Model Over Context Sensitive Token Encodings.* ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1. https://doi.org/10.18653/v1/p18-1246, (2018).

[23] Khoe, Y. H., *Reproducing a Morphosyntactic Tagger With a Meta-Bilstm Model Over Context Sensitive Token Encodings.* LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings. (2020).

[24] Elov, B., Axmedova, X., *Determining Homonymy Using Statistical Methods. Computational Models and Technologies*, Uzbekistan-Malaysia international conference, September 16-17th, 2022, Tashkent. (2022).

[25] Elov, B., Axmedova, X., *Business Process Modeling That Distinguishes Homonymy Within Three Parts Of Speechs in Uzbek Language.* 7th International Conference on Computer Science and Engieering, 14-16 September 2022, Turkey, (2022).

[26] Zucchini, W., MacDonald, I.L., Langrock, R. (2016). *Hidden Markov Models for Time Series:* An Introduction Using R, Second Edition (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b20790

[27] Al-Anzi, F. ve AbuZeina, D., *A Survey of Markov Chain Models in Linguistics Applications.* 53-62. 10.5121/csit.2016.61305. (2016).

[28] Kadim, A. ve Lazrek, A., *Bidirectional HMM-based Arabic POS tagging.* International Journal of Speech Technology. 19. 10.1007/s10772-015-9303-7. (2016).