یک مدل بیزی برای استخراج با مربی گرامر زبان طبیعی

هشام فیلی[1]، حمیدرضا قادر[2] و مرتضی آنالویی[3]
[1 و 2] آزمایشگاه پردازش زبان طبیعی، دانشکدهٔ برق و کامپیوتر، دانشگاه تهران
[3] گروه هوش مصنوعی و رباتیک، دانشکدهٔ کامپیوتر، دانشگاه علم و صنعت

چکیده

در این مقاله نشان داده‌ایم که مسئلهٔ استخراج با مربی گرامر برای گرامرهای زبان طبیعی، می‌تواند به‌عنوان ترکیبی پیچیده از تعداد زیادی مسئله انتخاب مدل تعریف شود. مسائل انتخاب مدل به مسائلی گفته می‌شود که در آنها میان مدل‌هایی با پیچیدگی آماری متفاوت تصمیم‌گیری می‌شود. برای ارائهٔ مدل مورد نظر، ابتدا یک مدل بیزی از فرایند شناختی انتخاب مدل را معرفی کرده‌ایم. این مدل مسئلهٔ انتخاب مدل را با تطابق بیشتر با رفتار انسان حل می‌کند. سپس با استفاده از تعمیم‌یافتهٔ مدل بیزی، که فرایند دیریکله است، به‌همراه مفهوم گرامرهای مبتنی بر سابقه، یک مدل بیزی مبتنی‌بر سابقه برای استخراج با مربی گرامر ارائه کرده‌ایم. نتایج آزمایشی که با این مدل روی یک پیکرهٔ استاندارد برای زبان انگلیسی صورت گرفته است، در مقایسه با مدل مرجع طبق معیار F1 به مقدار ٪۹/۱ پیشرفت نشان می‌دهد.

واژگان کلیدی: مدل بیزی، گرامر جایگزینی درخت، فرایند دیریکله، فرایند رستوران چینی، گرامرهای مبتنی بر سابقه.

۱- مقدمه و کارهای پیشین

مجموعهٔ تحقیقات انجام شده روی استخراج گرامر، در دو دههٔ اخیر، حاکی از اهمیت گرامر محاسباتی در پردازش زبان طبیعی است *(Cohn, et al., 2009; Johnson, 1998; Collins, 1999; Charniak, 2000; Charniak, et al., 2005)*. در استخراج گرامر به‌صورت بامربی[1]، گرامر محاسباتی از روی درخت‌های تجزیهٔ جملات که از قبل توسط زبان‌شناسان تهیه شده است، استخراج می‌شود. بدین معنا که جملات یک پیکره توسط انسان، تجزیهٔ نحوی شده و درخت‌های تجزیهٔ[2] این جملات تهیه می‌شود و مجموعهٔ این درخت‌ها با نام بانک درختی[3] برای استخراج قواعد پایهٔ گرامر مورد استفاده قرار می‌گیرد. تلاش‌های زیادی در گذشته برای استخراج گرامرهای مستقل از متن احتمالاتی[4] صورت گرفته است تا گرامرهایی استخراج شود که نتایج خوبی در تجزیهٔ نحوی جملات زبان طبیعی تولید کنند *(Cohn, et al., 2009)*. این تلاش‌ها شامل ایجاد و به‌کارگیری روش‌هایی برای تقویت صورت‌گرای[5] توصیف کنندهٔ گرامر (*Johnson, 1998*) (*Collins, 1999*)، روش‌های مسطح‌سازی پیشرفته (*Charniak, 2000*) *(Charniak, et al., 2005)* و روش‌های انتخاب ویژگی (*Charniak, 2000*) است. تمام این تلاش‌ها که در کار ارائه شده توسط کُهن و همکاران *(Cohn, et al., 2009)* تحت عنوان مهندسی گرامر از آنها یاد شده، برای پوشش ضعف‌های گرامرهای مستقل از متن در توصیف جملات زبان طبیعی است. بیشتر ضعف‌های گرامرهای مستقل از متن در توصیف زبان‌های طبیعی ناشی از فرض‌های استقلالی است که در این گرامرها می‌شود. به‌عنوان مثال، برخی خواص زبان‌های طبیعی نظیر خاصیت حذف ضمیر در جایگاه فاعل

[1] Supervised
[2] Parse Trees
[3] Tree-bank
[4] Probabilistic Context Free Grammars
[5] Formalism

در زبان فارسی (فیلی, ۱۳۸۵) و خاصیت وابستگی Cross-Serial در زبان‌های آلمانی، سوییسی و هلندی *(Mohri, et al., 2006)* خواص حساس به متن هستند و گرامرهای مستقل از متن توانایی توصیف آنها را ندارند. در مدل‌هایی که مورد اشاره شد، با تغییر بانک درخت یا زبان تمامی این تلاش‌ها برای بالا بردن دقت مدل باید دوباره انجام گیرد *(Cohn, et al., 2009)*. در این مقاله، ما مدلی بامربی را ارائه می‌کنیم که گرامر جایگزینی درخت احتمالاتی[1]، که صورت‌گرایی قوی‌تر از گرامرهای مستقل از متن در توصیف زبان طبیعی است، استخراج می‌کند. علاوه‌بر این، مدل ما از دیدگاه آماری در قالب روش‌های بیزی غیرپارامتری[2] دسته‌بندی می‌شود. در این روش‌ها تعداد پارامترهای مدل محدود فرض نمی‌شود، از این‌رو پیچیدگی مدل با پیچیدگی داده‌های آموزشی به‌طور خودکار تنظیم شده و مشکلات برازش کم یا زیاد[3] پیش نمی‌آید *(Teh, 2010)*. بدین‌ترتیب مدل‌های بیزی غیرپارامتری پیچیدگی‌های داده‌ی آموزشی را به‌خوبی فرامی‌گیرند. این خاصیت به‌همراه قوی‌تر بودن صورت‌گرای استخراجی در مدل ما سبب می‌شود، بسیاری از پیچیدگی‌های زبانی که در مدل‌های پارامتری با روش‌های مهندسی گرامر در مدل کد می‌شد، به‌طور خودکار از داده‌های آموزشی فراگرفته شود.

یکی از اوّلین کارهای شاخص در زمینهٔ استخراج گرامر، که به استخراج گرامر مستقل از متن احتمالاتی پرداخته است، کار ارائه شده توسط چارنیاک *(Charniak, 1996)* است. در این کار با استفاده از یک مدل مبتنی‌بر درست‌نمایی[4] بیشینهٔ گرامری با ۱۰۶۰۵ قانون استخراج شد. یک برتری شاخص این کار نسبت به مدل‌های موجود در آن زمان، تجزیه‌گر ارائه شده در این کار بود که توانست با استفاده از گرامر ایجاد شده، بالاترین دقّت تجزیهٔ جملات در آن زمان را به دست آورد. در این کار برای تجزیهٔ جملات با طول کمتر از یک‌صد کلمهٔ موجود در پیکرهٔ WSJ[5] (WSJ100) مقدار ٪۷۹/۵ برای معیار *F1* گزارش شده است. پس از آن تلاش‌هایی برای توسعه مدل‌های گرامرهای مستقل از متن آغاز شد تا ضعف‌های این گرامرها را در تجزیهٔ جملات پوشش دهند. در کار ارائه شده توسط جانسون *(Johnson, 1998)* روشی برای استخراج گرامرهای مستقل از متن، مبتنی‌بر سابقه[6] ارائه شد. در *(Collins, 1999)* کار دیگری ارائه شد که با ایجاد وابستگی واژگانی[7] توانست صورت‌گرا را تقویت کند. این کار تجزیه‌گر مناسب برای این نوع گرامر را نیز ارائه کرد و توانست برای تجزیهٔ جملات WSJ100 مقدار ٪۸۸/۲ را برای معیار *F1* به‌دست آورد. چارنیاک[8] توانست در *(Charniak, 2000)* با استفاده از یک تجزیه‌گر و روش مسطح‌سازی[9] جدید، که هر دو بر پایهٔ آنتروپی بیشینه[10] بودند، مقدار ٪۸۹/۵ را برای جملات موجود در WSJ100 به‌دست آورد. این کار در *(Charniak, et al., 2005)* با استفاده از یک الگوریتم رتبه‌بندی[11] تقویت شد و مقدار *F1* را برای جملات WSJ100 به ٪۹۱/۰ رساند. در *(Petrov, et al., 2007)* روشی برای استخراج گرامرهای مستقل از متن احتمالاتی ارائه شد که با وجود عدم وابستگی واژگانی در گرامر استخراجی، توانست دقتی قابل رقابت با بهترین گرامرهای دارای وابستگی واژگانی به‌دست آورد. در مدل ارائه شده در این کار، در یک فرایند تکراری، هر غیرپایانه به دو غیرپایانه تقسیم می‌شد تا بدین‌ترتیب بتواند تخمین بهتری از پارامترهای آزاد مدل ایجاد کند *(Cahill, 2008)*. مقدار F1 گزارش شده برای جملات موجود در WSJ100 در این کار برابر ٪۹۰/۰ است. براساس دانش نگارندگان، بهترین نتایج تجزیهٔ نحوی مربوط به کار ارائه شده در *(McClosky, et al., 2006)* است. در این کار با آموزش تجزیه‌گر نحوی روی نتایج تجزیهٔ خود دقت عملکرد تجزیه‌گر افزایش داده شده است. در این کار مقدار گزارش شده برای معیار F1 در تجزیهٔ جملات موجود در WSJ100 برابر ٪۹۲/۱ است.

مدل *DOP (Bod, 1992)* نیز مدل موفقی است که گرامر استخراجی آن در دستهٔ گرامرهای جایگزینی درخت احتمالاتی قرار می‌گیرد. البته این مدل به‌طور واضح گرامر استخراج نمی‌کند، بلکه با استفاده از داده‌های آموزشی و زیردرخت‌های پرتکرار در این داده‌ها، بهترین تجزیهٔ نحوی برای جملهٔ ورودی را محاسبه می‌کند. این مدل در *(Bod, 2003)* به مقدار ٪۹۰/۷ برای معیار *F1* در تجزیهٔ جملات WSJ100 دست یافت.

مدل‌های بیزی غیرپارامتری در بسیاری از

[1] Probabilistic Tree Substitution Grammar
[2] Non-parametric Bayesian Models
[3] Under-fit or over-fit
[4] Maximum Likelihood
[5] Wall Street Journal

[6] History-based Grammars
[7] Lexicalization
[8] Charniak
[9] Smoothing
[10] Maximum Entropy
[11] Reranking

کاربردهای پردازش زبان طبیعی به‌کار گرفته شده‌اند. در *(Goldwater, et al., 2007)* مسئلهٔ بدون‌ناظر برچسب‌گذاری جزء کلام با استفاده از این مدل‌ها حل شده است. در *(Johnson, et al., 2007)* قالبی[1] برای یادگیری لغات از دادهٔ گویشی ارائه شده است، که بر پایهٔ مدل‌های بیزی غیرپارامتری بنا شده است. در *(Brody, et al., 2009)* حل مسئله ابهام‌زدایی معنای لغت[2] با استفاده از این مدل‌ها، مطرح شده است. برچسب‌گذاری نقش معنایی نیز در *(Titov, et al., 2011)* با این مدل‌ها حل شده است.

مدل‌های بیزی غیرپارامتری پیش از این برای استخراج بامربی گرامر جایگزینی درخت نیز در *(Cohn, et al., 2010)*، *(Cohn, et al., 2009)* و *(O'Donnell, et al., 2009)* به کار گرفته شده‌اند. در *(Cohn, et al., 2010)* و *(Cohn, et al., 2009)* فرایند دیریکله[3] و فرایند پیتمن-یور[4]، که شکل تعمیم‌یافتهٔ فرایند دیریکله است، برای استخراج گرامر جایگزینی درخت احتمالاتی به‌کار رفته‌اند. این دو فرایند نمونه‌هایی از مدل‌های بیزی غیر پارامتری هستند. در این دو کار، هر واحد پایه در گرامر به شرط گرهٔ ریشهٔ آن واحد، با استفاده از یکی از دو فرایند مذکور، استخراج می‌شود. بدین‌ترتیب مفهوم سابقه در این دو کار مد نظر قرار نگرفته است. مفهوم سابقه موجب وابستگی یادگیری هر واحد گرامر به بخشی از بافت جمله که آن واحد در آن رخ می‌دهد، می‌شود. این مفهوم در *(Johnson, 1998)* برای کاهش اثر سوءفرض‌های استقلال در استخراج گرامرهای مستقل از متن احتمالاتی مطرح شد. در *(Johnson, 1998)* برای استخراج گرامرهای مبتنی‌بر سابقه، روشی ارائه شد که براساس آن غیرپایانه‌های درخت‌های تجزیهٔ موجود در داده‌های آموزشی، با برچسب غیرپایانهٔ پدر خود برچسب‌گذاری می‌شوند. بدین‌ترتیب واحد زبانی رخدادیافته در زیر یک غیرپایانه، به بخشی از بافت جمله که زیر غیرپایانه‌ی پدر ریشه‌اش قرار دارد، وابسته می‌شود و این وابستگی در فرایند یادگیری، تأثیر می‌گذارد. علاوه‌بر این، در این دو کار مبنای شناختی فرایندهای دیریکله و پیتمن-یور مورد اشاره قرار نگرفته است و علّت به‌کارگیری تعداد زیادی فرایند دیریکله و پیتمن-یور محدودیتی بیان شده است که در گرامر استخراجی از یک فرایند، بروز می‌کند. امّا در مدل ارائه شده در این مقاله، ما مبنای شناختی فرایند دیریکله را به‌کار گرفته و نشان می‌دهیم که چگونه این مدل مبنایی، منجر به تعداد زیادی فرایند دیریکله در مسئلهٔ استخراج گرامر می‌شود. سپس مفهوم سابقه را نیز در مدل خود وارد کرده و تغییرات این مفهوم بر فرایند دیریکله را نشان می‌دهیم.

مدل ارائه شده در *(Johnson, et al., 2007)* یک نمونهٔ دیگر برای استخراج گرامر با مدل‌های بیزی غیر پارامتری است. در این کار یک قالب برای استخراج گرامر با مدل‌های بیزی غیرپارامتری ارائه می‌شود که آن را گرامر تطبیقی[5] می‌نامند. این مدل برای یادگیری لغات از دادهٔ گویشی در *(Johnson, et al., 2009)* به‌کار گرفته شده است. اما این گرامر برای توصیف جملات زبان طبیعی مناسب نیست؛ چرا که بر پایهٔ گرامرهای مستقل از متن بنا شده است.

در *(O'Donnell, et al., 2009)* نیز فرایند پیتمن-یور برای استخراج گرامر جایگزینی درخت به‌کار رفته است. در این کار، استخراج گرامر با مدل‌های بیزی غیرپارامتری از دیدگاه شناختی مدل شده و مورد بررسی و آزمایش قرار گرفته است. کار مذکور یک قالب بیزی غیرپارامتری برای استخراج گرامرهایی که توانایی خوبی در توصیف جملات زبان طبیعی دارند، ارائه می‌کند. درحقیقت قالب ارائه شده در این کار تعمیم‌یافتهٔ گرامر تطبیقی است. نتیجهٔ مدل ارائه شده در این کار نیز، همانند مدل ارائه شده در *(Cohn, et al., 2009)* و *(Cohn, et al., 2010)*، تعداد زیادی فرایند دیریکله برای یادگیری واحدهای زبانی به شرط گرهٔ ریشه است. همچنین، صورت‌گرای استخراجی در این کار نیز گرامر جایگزینی درخت احتمالاتی است. علاوه‌بر این، در این کار نیز همانند کارهای ارائه شده در *(Cohn, et al., 2010)* و *(Cohn, et al., 2009)* مفهوم سابقه در استخراج گرامر دخیل نمی‌شود.

بنابر آنچه گفته شد، در این مقاله یک مدل بامربی، بر پایهٔ روش‌های بیزی غیرپارامتری، برای استخراج گرامر زبان طبیعی ارائه شده است. صورت‌گرای گرامر استخراج شده در این مقاله یک مدل قوی‌تر از گرامر مستقل از متن است که گرامر جایگزینی درخت احتمالاتی نامیده می‌شود. کار این مقاله به کار ارائه شده در *(Cohn, et al., 2009)* شباهت دارد و درواقع توسعه‌یافتهٔ آن به‌شمار می‌رود. نوآوری‌های این کار نسبت به کارهایی که پیش از این مدل‌های مشابهی برای استخراج گرامر ارائه کرده‌اند، دو

[1] Framework
[2] Word Sense Disambiguation
[3] Dirichlet Process
[4] Pitman-Yor Process
[5] Adaptor Grammar

بخش دارد. اوّل اینکه ما نشان می‌دهیم چگونه مدل ارائه‌شده برای استخراج گرامر از یک مدل بیزی ساده‌تر، که با عنوان مدل بیزی از فرایند شناختی انتخاب مدل[1] معرّفی می‌کنیم، ناشی می‌شود. در این میان، دلیل اینکه تعداد زیادی فرایند دیریکله در مدل ارائه‌شده برای استخراج گرامر به‌وجود می‌آید، روشن می‌شود. نتیجهٔ این بخش از مدل‌سازی ما مشابه مدل‌های ارائه‌شده در *(Cohn, et al., 2009; Cohn, et al., 2010; O'Donnell, et al., 2009)* می‌گردد. دوم اینکه، ما مدل ارائه‌شدهٔ خود را با دخیل‌کردن مفهوم سابقه طراحی می‌کنیم و نشان می‌دهیم که این توسعه منجر به نتایجی بهتر از نتایج ارائه شده در *(Cohn, et al., 2010)* و *(Cohn, et al., 2009)* می‌شود. روشی که در این مقاله معرّفی شده روی زبان انگلیسی مورد آزمایش قرار گرفته است. نتایج ارزیابی روی بخش WSJ پیکرهٔ استاندارد Penn Treebank، که به‌طور تقریبی تمامی کارهای معتبر در زبان انگلیسی نیز روی آن تعریف شده‌اند، نشان‌دهندهٔ برتری روش ارائه‌شده در این مقاله نسبت به کارهای مشابه است.

در ادامه، در بخش دو پیش‌زمینهٔ تئوری را که برای درک مدل ارائه‌شده در این مقاله لازم است معرّفی، سپس در بخش سه مدل خود را ارائه و در بخش چهار چگونگی استنتاج در مدل ارائه‌شده را بیان می‌کنیم. پس از آن در بخش پنج آزمایش‌های انجام شده را توضیح داده و نتایج حاصل را مورد تحلیل قرار می‌دهیم. در بخش پایانی نیز نتیجه‌گیری خود را از مدل ارائه‌شده و آزمایش‌ها بیان می‌کنیم.

۲- پیش‌زمینه

روش‌های بیزی غیر پارامتری و استنتاج بیزی[2] با استفاده از روش‌های مونت کارلوی زنجیرهٔ مارکوف[3]، روش‌هایی هستند که در نیمهٔ دوم دههٔ گذشتهٔ میلادی در حلّ مسائل حوزهٔ پردازش زبان طبیعی به‌کار گرفته شده‌اند *(Goldwater, et al., 2007; Goldwater, et al., 2009; Cohn, et al., 2009; Cohn, et al., 2010; Johnson, et al., 2007)*. در این بخش، ابتدا به معرفی مدل بیزی از فرایند شناختی انتخاب مدل، که مبنای تئوری مدل ارائه شده در این مقاله را تشکیل می‌دهد می‌پردازیم. سپس فرایند دیریکله را که یک روش بیزی غیرپارامتری است و تعمیم بی‌نهایت مدل بیزی شناختی از فرایند انتخاب مدل است، معرّفی کرده و پس از آن روش نمونه‌برداری گیبس[4] را، که یک روش مونت کارلوی زنجیرهٔ مارکوف است، توضیح می‌دهیم. در این مقاله، از نمونه‌برداری گیبس برای تخمین توزیع احتمال حاصل از فرایندهای دیریکله استفاده می‌کنیم. در پایان نیز اشاره‌ای به‌طور تقریبی به گرامر جایگزینی درخت می‌کنیم.

۲-۱- مدل بیزی از فرایند شناختی انتخاب مدل

مدل بیزی‌ای را که مبنای مدل ارائه‌شده در این تحقیق تشکیل می‌دهد یک مدل بیزی از فرایند شناختی انتخاب مدل است. بسیاری از مسائل در دنیای واقعی هستند که در آنها باید میان دو یا چند فرضیه که از نظر پیچیدگی با هم متفاوت هستند و فرضیه‌های غیر همگن به‌شمار می‌آیند، تصمیم‌گیری کرد. این‌گونه مسائل که در آنها دانش حاصل از داده‌های مشاهده شده را برای انتخاب میان دو مدل آماری که از نظر پیچیدگی با هم متفاوت هستند به‌کار می‌گیریم، *انتخاب مدل* نامیده می‌شوند *(Griffiths, et al., 2008)*.

برای درک عملکرد مدل بیزی شناختی از فرایند انتخاب مدل یک مثال استاندارد را ارائه می‌کنیم. این مثال از *(Griffiths, et al., 2008)* برگرفته شده است. فرض کنید که یک سکه داریم و θ نشان‌دهنده‌ی احتمال رخداد شیر در پرتاب آن است. می‌خواهیم تصمیم بگیریم که این سکه، یک سکهٔ معمولی با $\theta = \frac{1}{2}$ است یا سکه‌ای است که در آن θ مقدار خود را از یک توزیع یکنواخت بین ۰ و ۱ گرفته است. این فرضیه به مفهوم آن است که سکهٔ معمولی نیست و احتمال شیر و خط آمدن، لزوماً مساوی نیست؛ ولی اینکه مقدار احتمال چقدر باشد در این فرضیه مهم نیست. مهم آن است که بدانیم سکه اریب است. بنابراین دو فرضیه وجود دارد:

۱. $\boldsymbol{\theta = \frac{1}{2}}$ که آنرا h0 می‌نامیم.
۲. θ مقدار خود را از یک توزیع یکنواخت بین ۰ و ۱ گرفته است. این فرضیه را h1 می‌نامیم.

فرض می‌کنیم که هیچ دانشی مبنی‌بر برتری داشتن یکی از فرضیات بر دیگری وجود ندارد. بنابراین می‌توان فرض کرد که $\mathbf{P(h_0) = P(h_1) = \frac{1}{2}}$. حال فرض کنید که ما یک ترتیب ده تایی را از پرتاب این سکه مشاهده می‌کنیم،

[1] Model Selection
[4] Gibbs Sampling
[4] Gibbs Sampling

[4] Gibbs Sampling

به‌عنوان مثال {ش،ش،خ،ش،خ،خ،ش،ش،ش،خ}. اگر ما از روش درست‌نمایی بیشینه[1] (MLE) استفاده کنیم برای مشاهدهٔ ترتیب فوق، خواهیم داشت $\hat{\theta} = N_{ش}/10$ که در آن $N_{ش}$ نشان‌دهندهٔ تعداد رخداد شیر است. در نتیجهٔ کاربرد این روش، $P(h|w,\hat{\theta})$ برای هر ترتیب مشاهده شده‌ای که تعداد نامساوی از رخداد شیر و خط دارد، همواره به h_1 احتمال بیشتری اختصاص می‌دهد. همچنین برای هر ترتیبی که رخداد برابر شیر یا خط دارد، احتمال h_1 و h_0 برابر می‌شود. بدین‌ترتیب چنین راه حلی همواره فرض h_1 را برای جواب این مسئله انتخاب می‌کند که این موضوع به‌طور کامل مخالف رفتاری است که انسان در حل چنین مسئله‌ای از خود نشان می‌دهد. حال اگر از مدل بیزی مورد نظر استفاده کنیم، حل مسئله به‌گونه‌ای دیگر خواهد بود. در این حالت ما توزیع احتمال روی پاسخ مسئله را با انتگرال‌گیری روی مقادیر ممکن برای θ پیدا می‌کنیم. یعنی:

$$P(h|w) = \int P(h|w,\theta)P(\theta|w)d\theta \qquad (۱)$$

رابطهٔ فوق منجر به مقدار احتمال زیر خواهد شد.

$$P(h_1|w) = 1/(1+\frac{11!}{N_{ش}!\,N_{خ}!\,2^{10}}) \qquad (۲)$$

این رابطه زمانی که تعداد رخداد شیر و خط برابر باشد بسیار کوچک‌تر از ۰.۵ خواهد بود و فقط زمانی h_1 را بر h_0 برتری می‌دهد که $N_{ش} \geq 8$ یا $N_{ش} \leq 2$ باشد. این نتیجه با رفتاری که از انسان در حل چنین مسئله‌ای مشاهده می‌شود مطابقت بیشتری دارد و درحقیقت مدل مناسب‌تری از فرایند شناختی انتخاب مدل در انسان است.

آنچه در حل این مسئله به روش بیز پنهان است بدین‌ترتیب بیان می‌شود:

مدل‌ها و فرضیات آماری پیچیده‌تر دارای درجهٔ آزادی بیشتری هستند. به همین دلیل همواره می‌توان آنها را بهتر از مدل‌ها و فرضیات ساده بر داده‌های مشاهده‌شده منطبق کرد؛ به‌گونه‌ای که داده‌های مشاهده‌شده تحت مدل پیچیده‌تر احتمال تولید بیشتری داشته باشد *(Griffiths, et al., 2008)* به‌عنوان مثال، در مثال پرتاب سکه، برای هر ترتیبی از شیر و خط می‌توان مقداری برای θ یافت که احتمال بیشتری به تولید آن ترتیب خاص نسبت به $\theta = \frac{1}{2}$ بدهد. بدین‌ترتیب به نظر می‌رسد که همواره باید در این‌گونه مسائل مدل پیچیده‌تر را بر مدل ساده برتری داد. امّا نگرش مدل بیزی در حلّ این‌گونه مسائل، این موضوع را دخالت می‌دهد که هر چند درجهٔ آزادی بیشتر امکان تطبیق بیش‌تر مدل با داده‌های مشاهده شده را فراهم می‌کند، امّا این درجهٔ آزادی امکان یک تطبیق بدتر را نیز ایجاد می‌کند. نگرش مدل بیزی در حل این‌گونه مسائل با انتگرال‌گیری روی درجات آزادی مدل پیچیده‌تر، میان حالاتی که مدل پیچیده‌تر انطباق بهتری در تولید داده ایجاد می‌کند و حالاتی که این مدل انطباق بدتری نسبت به مدل ساده، در تولید داده‌ها ایجاد می‌کند، میانگین می‌گیرد *(Griffiths, et al., 2008)*. بنابراین نگرش مدل بیزی، مدل پیچیده‌تر را بر اساس انطباق بیشتر مدل، زمانی که پارامترهای مدل بهترین مقادیر را اختیار کرده‌اند، انتخاب نمی‌کند. بلکه زمانی مدل پیچیده‌تر را بر مدل ساده‌تر برتری می‌دهد که مدل پیچیده‌تر با انتخاب تصادفی مقادیر پارامترها، در بیشتر مواقع، انطباق بهتری بر روی داده‌های مشاهده شده ایجاد کند. این موضوع با قضاوت انسان در حل این مسائل مطابقت دارد، چرا که با فرض در نظر گرفتن مثال پرتاب سکه، انسان نیز، همواره و با مشاهدهٔ هر ترتیبی از شیر و خط، مدل پیچیده‌تر را به مدل ساده‌تر ترجیح نمی‌دهد؛ بلکه با فرض اینکه پارامترهای مدل پیچیده‌تر مقدار خود را به‌صورت تصادفی بگیرند، قضاوت می‌کند که آیا در بیشتر موارد مدل پیچیده‌تر انطباق بهتری روی داده‌ها تولید می‌کند یا مدل ساده‌تر.

۲-۲- فرایند دیریکله

فرایند دیریکله یک فرایند تصادفی است که درحقیقت به‌گونه‌ای نتیجهٔ مدل بیزی از فرایند شناختی انتخاب مدل برای مسائلی است که مدل‌های مورد تصمیم‌گیری دارای پیچیدگی نامعلوم هستند. برای به‌دست آورن درک بهتر از چنین مسائلی یک مثال می‌زنیم. با فرض در نظر داشتن مثال سکه‌ای که در بخش قبل مورد بحث قرار گرفت، فرض کنید ما ترتیبی از پرتاب یک جسم چندوجهی را مشاهده کرده‌ایم ولی تعداد وجوه جسم مورد نظر را نمی‌دانیم و می‌خواهیم همانند مثال سکه در مورد توزیع احتمال جسم مورد نظر تصمیم‌گیری کنیم. این یک مسئلهٔ انتخاب مدل بسیار پیچیده است.

فرایند دیریکله یک فرایند تصادفی است که در مدل‌های بیزی غیرپارامتری به‌کار گرفته می‌شود. این فرایند تصادفی درحقیقت یک توزیع احتمال روی توزیع

[1] Maximum likelihood

احتمال‌هاست، یعنی یک رخداد نمونه‌گیری‌شده از این توزیع احتمال، خود یک توزیع احتمال است. توزیع احتمال‌های نمونه‌گیری شده از فرایند دیریکله توزیع احتمال‌های ناپیوسته (توزیع احتمال جسم n وجهی، که n می‌تواند بی‌نهایت باشد) هستند. این توزیع‌های احتمال را نمی‌توان با تعداد محدودی از پارامترها توصیف کرد و از این‌رو فرایند دیریکله در دستهٔ مدل‌های غیرپارامتری قرار می‌گیرد (*Teh, 2010*). درحقیقت فرایند دیریکله توزیع احتمالی است که دامنهٔ آن دنیای توزیع‌های احتمال روی یک متغیر تصادفی است.

گفتیم دامنهٔ یک فرایند دیریکله توزیع‌های احتمال روی یک متغیر تصادفی است، بنابراین می‌توان یک توزیع احتمال نمونه‌گیری شده را از یک فرایند دیریکله یک توزیع احتمال تصادفی به حساب آورد.

فرض کنید که G یک توزیع احتمال تصادفی است؛ چیزی همانند یک متغیر تصادفی، که دارای توزیع فرایند دیریکله است. برای اینکه این چنین باشد، باید توزیع‌های حاشیه‌ای[1] روی این توزیع احتمال تصادفی دارای توزیع دیریکله[2] باشند (*Teh, 2010*). برای توضیح این موضوع، فرض کنید که H یک توزیع روی متغیر تصادفی پیوسته Θ است و α یک عدد حقیقی مثبت است. برای هر افراز متناهی روی Θ که به‌صورت $\mathbf{A_1}, \dots, \mathbf{A_r}$ نشان می‌دهیم، بردار $[\mathbf{G(A_1)}, \dots, \mathbf{G(A_r)}]$ یک بردار تصادفی است، چرا که G یک توزیع احتمال تصادفی است. حال می‌توانیم بگوییم که G یک توزیع احتمال تصادفی توزیع شده بر حسب فرایند دیریکله با توزیع مبنای H است، اگر داشته باشیم:

$$[\mathrm{G(A_1)}, \dots, \mathrm{G(A_r)}] \sim \mathrm{Dirichlet}(\alpha \mathrm{H(A_1)}, \dots, \alpha \mathrm{H(A_r)}) \qquad (۳)$$

که برای هر افراز متناهی $(\mathbf{A_1}, \dots, \mathbf{A_r})$ از Θ برقرار باشد (*Teh, 2010*).

این یعنی توزیع‌های حاشیه‌ای روی توزیع G دارای توزیع دیریکله باشند. در این صورت می‌گوییم G دارای توزیع فرایند دیریکله است و نشان می‌دهیم:

$$\mathrm{G} \sim \mathrm{DP}(\alpha, \mathrm{H}) \qquad (۴)$$

حال فرض کنید که $\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_n}$ یک ترتیب نمونه‌گیری شده مستقل با توزیع یکسان (i.i.d[3]) از G است.

برای توزیع احتمال $\mathbf{P(\theta_{n+1} | \theta_1, \dots, \theta_n)}$ با انتگرال‌گیری از $\mathbf{P(G, \theta_{n+1} \in A \,|\theta_1, \dots, \theta_n)}$ روی G، نتیجه می‌شود:

$$P(\theta_{n+1} \in A | \theta_1, \dots, \theta_n) = \frac{1}{\alpha+n}\left(\alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A)\right) \qquad (۵)$$

در این جا G حکم متغیر نهانی را دارد که در بخش قبل نیز روی آن انتگرال گرفتیم. در رابطهٔ فوق $A \subset \Theta$ است و $\delta_{\theta_i}(A)$ تابع دلتای کرونکر[4] است که تمام احتمال خود را در نقطه θ_i متمرکز کرده است (*Teh, 2010*). بنابراین داریم:

$$\theta_{n+1} | \theta_1, \dots, \theta_n \sim \frac{1}{\alpha+n}\left(\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}\right) \qquad (۶)$$

برای درک این رابطه فرض کنید $\{\mathbf{\theta_1^*}, \dots, \mathbf{\theta_{|c|}^*}\}$ نشان‌دهندهٔ تمامی مقادیر مختلفی است که $\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_n}$ گرفته‌اند. حال فرض کنید $\mathbf{c} = \{\mathbf{c_1}, \dots, \mathbf{c_n}\}$ برداری از نمایه‌ها برای $\{\mathbf{\theta_1^*}, \dots, \mathbf{\theta_{|c|}^*}\}$ باشد، به‌طوری‌که $\boldsymbol{\theta_i} = \mathbf{\theta_{c_i}^*}$ باشد. در این صورت توزیع $\boldsymbol{\theta_{n+1}}$ به‌صورت زیر خواهد بود:

$$\theta_n = \begin{cases} \theta_i^* \;\; with\ prob. \;\; \frac{|\{j : c_j = i\}|}{n-1+\alpha} \\ \theta, \;\; \theta \sim H \;\; with\ prob. \;\; \frac{\alpha}{n-1+\alpha} \end{cases} \qquad (۷)$$

در رابطهٔ فوق $\mathbf{|\{j : c_j = i\}|}$ بیان‌گر تعداد دفعاتی است که $\mathbf{\theta_i^*}$ در $\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_n}$ رخ داده است. رابطهٔ فوق نشان می‌دهد که فرایند دیریکله اثر خوشه‌بندی[5] دارد. قسمت اوّل رابطهٔ فوق نشان می‌دهد که یک دادهٔ مشاهده شده جدید با احتمال $\frac{|\{j : c_j = i\}|}{n-1+\alpha}$ در یکی از خوشه‌های موجود قرار می‌گیرد یا با توجّه به قسمت دوم رابطه به‌احتمال $\frac{\alpha}{n-1+\alpha}$ منجر به ایجاد یک خوشهٔ جدید می‌شود. این خاصیت مبنای یک خاصیت شناختی در فرایند دیریکله نیز است. این خاصیت پیاده‌سازی حافظه به‌وسیلهٔ توزیع احتمال است.

انتگرال‌گیری از پارامتر آزاد فرایند دیریکله که موجب ایجاد رابطهٔ (۵) می‌شود، یک نگاه جدید را به فرایند دیریکله مطرح می‌کند، که به آن فرایند رستوران چینی[6] می‌گویند. در این نگاه، رابطهٔ (۵) نشان‌دهندهٔ توزیع احتمال انتخاب میزهای رستوران توسط مشتریانی است که به رستوران وارد می‌شوند. این توزیع احتمال با ورود هر مشتری، به‌گونه‌ای تغییر می‌کند که احتمال انتخاب میزهای شلوغ‌تر توسط مشتریان بعدی افزایش پیدا می‌کند. این خاصیت را خاصیت *"ثروتمند، ثروتمندتر می‌شود"*[7] می‌نامند.

[1] Marginal Distributions
[2] Dirichlet Distribution
[3] Independent and Identically distributed

[4] Kronecker Delta Function
[5] Clustring
[6] Chinese Restaurant Process
[7] Rich-get-richer

۲-۳- مونت کارلوی زنجیرهٔ مارکوف

استفاده از روش استنتاج بیزی برای یافتن توزیع احتمال پسین[1] در یک مسئلهٔ یادگیری، ممکن است منجر به یافتن توزیع احتمالی شود که به‌صورت تحلیلی قابل محاسبه نباشد. در این‌گونه موارد یک راه حل پرکاربرد استفاده از روش‌های مونت کارلوی زنجیره مارکوف برای تخمین تابع توزیع مورد نظر است. به‌عنوان مثال *(Cohn, et al., 2010)*، *(Johnson, et al., 2007 b)* و *(Goldwater, et al., 2007)* نمونه‌هایی از کاربرد این روش برای تخمین تابع توزیع پسین را در خود دارند. برای تخمین یک توزیع احتمال، روش‌های مونت کارلوی زنجیره مارکوف به خدمت گرفته می‌شوند تا تعداد زیادی نمونه از توزیع احتمال مورد نظر تولید کنند. این روش‌ها برای تولید نمونه از یک توزیع احتمال، یک زنجیره مارکوف می‌سازند که توزیع احتمال تولید شده توسط این زنجیره، معادل توزیع احتمال مورد نظر شود *(Johnson, et al., 2007 b)*. به بیان دیگر، این روش‌ها احتمال جابه‌جایی در زنجیرهٔ مارکوف را به‌گونه‌ای تعیین می‌کنند که احتمال مشاهدهٔ حالت[2] St_i در زنجیره، برابر با احتمال این حالت با توجّه به توزیع احتمال مورد نظر باشد. پس از تولید زنجیرهٔ مورد نظر، این روش‌ها با جابه‌جایی روی حالت‌های این زنجیره، تعداد زیادی نمونه از توزیع مورد نظر تولید می‌کنند *(Johnson, et al., 2007 b)*. البته زنجیرهٔ ایجاد شده به‌گونه‌ای است که توزیع احتمال آن به سمت توزیع احتمال مورد نظر میل می‌کند؛ بنابراین ابتدا لازم است تعداد زیادی نمونه از زنجیرهٔ مورد نظر تولیدشده و دور ریخته شود. بدین‌ترتیب پس از اینکه توزیع احتمال زنجیره به توزیع احتمال هدف به اندازهٔ کافی نزدیک شد، می‌توان نمونه‌های تولید شده را نمونه‌هایی از توزیع هدف تلقّی کرد. با تولید نمونه‌های زیاد از توزیع احتمال هدف می‌توان این نمونه‌ها را به طرق مختلف برای محاسبه مقادیر مورد نظر به‌کار گرفت. به‌عنوان مثال می‌توان این نمونه‌ها را برای تخمین MAP[3] توزیع احتمال هدف به‌کار گرفت *(Goldwater, et al., 2007)*. یک روش دیگر آن است که این نمونه‌ها را برای محاسبهٔ امید ریاضی روی تابع توزیع هدف به‌کار گرفت *(Johnson, et al., 2007 b)*. حال سعی می‌کنیم با روابط ریاضی ایدهٔ مبنایی این روش‌ها را شرح دهیم.

فرض کنید یک تابع داریم که با $\mathbf{f(z)}$ آن را نشان می‌دهیم. درصورتی‌که $\mathbf{z}$ یک متغیر پیوسته باشد رابطهٔ امید ریاضی تابع $\mathbf{f(z)}$ به‌صورت زیر بیان می‌شود:

$$E[f(z)] = \int f(z)P(z)dz \qquad (۸)$$

در این رابطه، $\mathbf{P(z)}$ توزیع احتمال پیشین روی $\mathbf{z}$ است. این انتگرال را می‌توان با استفاده از رابطهٔ:

$$E_{P(z)}[f(z)] \approx \frac{1}{N}\sum_{i=1}^{N} f(z_i) \qquad (۹)$$

به شرط اینکه $\mathbf{z_i}$ ها را از توزیع احتمال $\mathbf{P(z)}$ تولید کرده باشیم و تعداد آنها به اندازهٔ کافی زیاد باشد، تخمین زد. این نمونه‌ها را می‌توان از طریق زنجیرهٔ مارکوف تولید کرد.

تخمین فوق که در آن تعداد زیادی از نمونه‌های تولید شده از یک تابع برای تخمین آن به‌کار می‌رود، بیان‌گر بخش مونت کارلوی این روش‌ها است و تولید نمونه‌ها براساس توزیع احتمال $\mathbf{P(z)}$ به‌وسیلهٔ یک زنجیرهٔ مارکوف انجام می‌شود. از این‌رو، این روش‌ها را مونت کارلوی زنجیرهٔ مارکوف می‌نامند.

حال دقّت کنید که توزیع احتمال یک متغیر تصادفی را می‌توان به‌صورت توزیع حاشیه‌ای اشتراک آن متغیر با یک متغیر تصادفی دیگر نوشت. به‌عنوان مثال رابطهٔ زیر را در نظر بگیرید.

$$P(x) = \int P(x|\theta)P(\theta)d\theta \qquad (۱۰)$$

رابطهٔ فوق می‌تواند همانند رابطهٔ (۹) یک رابطه امید ریاضی محسوب شود. بنابراین می‌توان توزیع احتمال $\mathbf{P(x)}$ را همانند امید ریاضی $\mathbf{E_{P(z)}[f(z)]}$ تخمین زد. بنابراین از روش‌های مونت کارلوی زنجیرهٔ مارکوف می‌توان برای تخمین توزیع‌های احتمال بهره گرفت.

۲-۴- نمونه‌برداری گیبس

روش نمونه‌برداری گیبس یکی از روش‌های مونت کارلوی زنجیرهٔ مارکوف است که در مواقعی که توزیع احتمال مورد نظر، توزیع توأم تعداد زیادی متغیرهای تصادفی است، به‌کار گرفته می‌شود *(Resnik, et al., 2010)* . ایدهٔ مبنایی این روش آن است که هر یک از متغیرهای تصادفی را جداگانه و البته به شرط بقیهٔ متغیرهای تصادفی نمونه‌برداری کند *(Resnik, et al., 2010)*. این بدان معناست که از توزیع احتمالی نظیر آنچه در زیر می‌بینید، نمونه‌برداری کند.

$$P(rv_i^t | rv_1^t, \dots, rv_{i-1}^t, rv_{i+1}^{t-1}, \dots, rv_k^{t-1}) \qquad (۱۱)$$

[1] Posterior Distribution
[2] State
[3] Maximum a Posteriori

در این رابطه، $\mathbf{rv_i^t}$ نشان‌دهندهٔ متغیر تصادفی iام در مرحلهٔ tام است. همان‌گونه کـه از رابطـهٔ (۱۱) مشـخّص اسـت، هـر متغیر تصادفی در مرحلهٔ tام به آخـرین مقـادیر نسـبت داده شده به دیگر متغیرهای تصادفی وابسته است.

در این روش پس از اینکه از هر متغیر تصادفی یـک نمونه ایجاد شد، مجموعهٔ این نمونه‌ها در کنار هم یک نمونه از توزیع احتمال اشتراک متغیرها را تشـکیل مـی‌دهـد. کـلّ فرایند به این صورت است که به ازای هر متغیر تصادفی یک توزیـع چندجملـه‌ای محاسـبه مـی‌شـود و سـپس بـه روش Roulette Wheel یک مقدار برای متغیر تصادفی مورد نظـر، از توزیع محاسبه شده نمونه‌برداری می‌شود. این کار بـه ازای تمامی متغیرهای تصادفی تکرار می‌شود تا به‌ازای هر متغیـر تصادفی یک نمونه ایجاد شود. با تکرار این عملیات بـه‌انـدازهٔ کافی، می‌توان نمونه‌های زیادی از توزیع احتمال هدف تولید کرد که برای محاسبهٔ مقادیر مختلـف ماننـد امیـد ریاضـی و MAP توزیع احتمال هدف، به کار می‌روند.

برای یافتن MAP یک توزیـع احتمـال، مـی‌تـوان از روش‌های جستجوی نقاط بهینـهٔ توابـع اسـتفاده کـرد. یـک روش معمـول بـرای محاسـبهٔ MAP یـک توزیـع احتمـال، همراه‌سازی فرایند نمونه‌گیری با الگوریتم شـبیه‌سـازی ذوب فلز[1] است. درحقیقت، ایـن کـار منجـر بـه یـک جسـتجوی تپّه‌نوردی تصادفی[3] می‌شود که برای یـافتن نقطـه بیشـینه تـابع توزیـع احتمـال هـدف بـه‌کـار مـی‌رود. ایـن روش در *(Goldwater, et al., 2007)*، *(Cohn, et al., 2010)* بـه‌کـار گرفته شده است. این کار با الگـوریتم نمونـه‌بـرداری گیـبس به‌راحتی قابل انجام است. برای این منظور یک متغیر دمـا T تعریف می‌کننـد. سـپس در هنگـام نمونـه‌بـرداری از توزیـع احتمال هر متغیر تصادفی به شرط متغیرهای تصادفی دیگر، این توزیع احتمال را به توان $1/T$ می‌رسانند. متغیـر دمـا در هر دور نمونه‌برداری از تمامی متغیرهای تصادفی، ثابت نگـه داشته می‌شود. امّا در دور بعدی با یک گام مشخّص، کـاهش پیدا می‌کند. این کار را تا جایی ادامه می‌دهند که متغیر دما به صفر برسد. اثر این کار آن است که در الگوریتم تپّه‌نوردی، در مراحل اوّلیه، حالت‌های بدتر شانس بیشتری بـرای تولیـد می‌یابند و در ادامه این شانس کم‌تـر و کـم‌تـر مـی‌شـود. در *(Goldwater, et al., 2007)* اشاره شده اسـت کـه اگـر ایـن تغییر دما به اندازهٔ کافی کند صورت گیرد، تضمین می‌شـود که نمونه‌های آخر تولید شده از توزیع احتمـال هـدف حـول MAP این توزیع احتمال متمرکز شده‌اند.

در بخـش بعـد بـه تعریـف رسـمی صـورت‌گـرای استخراجی در این کار خـواهیم پرداخـت و بـه‌طـور مختصـر نکاتی را در مورد تئوری ایـن صـورت‌گـرا مـورد اشـاره قـرار خواهیم داد.

۲-۵- گرامر جایگزینی درخت احتمالاتی

همان‌طور که گفته شد، صورت‌گرای استخراجی در ایـن کـار گرامر جایگزینی درخت احتمـالاتی اسـت. ایـن صـورت‌گـرا قـوی‌تـر از گرامـر مسـتقل از مـتن احتمـالاتی در توصـیف زبان‌های طبیعی است. در این بخش، تعریف این صورت‌گرا و نحوهٔ اشتقاق در آن را توضیح مـی‌دهـیم تـا زمینـه را بـرای توصیف مدل مورد نظر برای استخراج این صورت‌گـرا فـراهم کنیم.

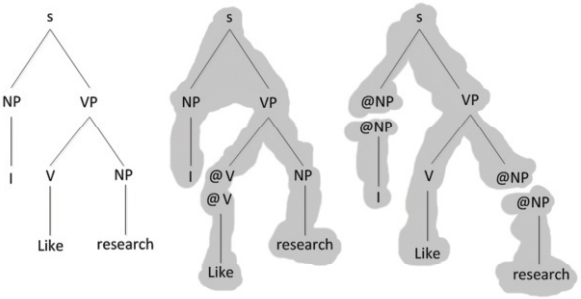
یک گرامر جایگزینی درخت احتمالاتی به‌صورت یک پنج تایی تعریف می‌شود *(Bod, 1995)*:

$$G =< V_N, V_T, R, P, \sigma >$$

که در آن $\mathbf{V_N}$ مجموعه نمادهـای غیـر پایـانی، $\mathbf{V_T}$ مجموعـه نمادهای پایانی، $\boldsymbol{\sigma} \in \mathbf{V_N}$ غیرپایانه شروع و $\mathbf{R}$ مجموعـه‌ای از قواعد تولید هستند که هر یک از این قوانین به‌صـورت یـک درخت هستند و به آنها درخت اوّلیه گفته می‌شود.$\mathbf{P}$ تـابعی اسـت کـه بـه هـر درخـت اوّلیـه $\mathbf{t \in R}$ یـک احتمـال $\mathbf{P(t)}$ تخصیص می‌دهد. برای یک درخـت t بـا ریشـه α، از $\mathbf{P(t)}$ به‌عنوان احتمال جانشینی درخـت t در غیـر پایانـه α تعبیـر می‌شود.

یک درخت اوّلیه، قطعه‌درختی به عمـق ≥ 2 اسـت که هر گره داخلی آن یک غیرپایانه بوده و گره‌های بـرگ در آن پایانه یا غیرپایانه هستند. برگ‌های غیرپایانه نقش مولد را در فرآیند تولید درختان در ایـن گرامـر ایفـا مـی‌کننـد. هـر اشتقاق، درختی را با شروع از ریشه تولید کـرده و هـر بـرگ غیرپایانه آن را با یک درخت اولیه جایگزین مـی‌کنـد و ایـن کار را تا زمانی ادامه می‌دهد که هیچ برگ غیرپایانه‌ای بـاقی نماند. برخلاف گرامـرهـای مسـتقل از مـتن، در ایـن گرامـر درخت‌های اشتقاق مختلف می‌تواننـد منجـر بـه تولیـد یـک درخت تجزیه شوند *(Bod, 1995)*. این موضوع در شکل زیـر نشان داده شده است. غیرپایانـه‌هـایی کـه بـا @ نشـان داده شده‌اند محل جایگزینی درخت اولیه هستند.

[1] Simulated Annealing
[3] Stochastic hill climbing



(شکل ۱): دو اشتقاق مختلف از یک درخت تجزیه. در سمت راست دو اشتقاق مختلف را که منجر به تولید درخت تجزیهٔ سمت چپ می‌شوند، مشاهده می‌کنید. گره‌های جایگزینی با @ علامت‌گذاری شده‌اند.

۳- مدل

یک پیکره از جملات تجزیه شده را در نظر بگیرید، آن‌را **E** می‌نامیم. فرض کنید **e** نیز نشان‌دهندهٔ ترتیب درخت‌های پایه‌ای است که درخت‌های پیکرهٔ **E** را تجزیه می‌کنند. ما در این مسئله می‌خواهیم بهترین **e** که پیکرهٔ مورد نظر را تجزیه می‌کند، پیدا کنیم. برای این منظور می‌توان توزیع احتمال **e** به شرط داده‌های مشاهده شده را یافت و از این توزیع احتمال استفاده کرد تا بهترین **e** که پیکرهٔ **E** را تجزیه می‌کند، به‌دست آوریم. به‌عنوان مثال، می‌توانیم MAP این توزیع احتمال را با استفاده از روش‌های موجود بیابیم. در ادامه راجع به این موضوع بیشتر توضیح خواهیم داد.

می‌توان نشان داد که مسئلهٔ استخراج گرامر جایگزینی درخت شبیه به مثال سکه‌ای است که در بخش ۲-۱ برای توصیف مدل بیزی از فرایند شناختی انتخاب مدل، ارائه شد و درحقیقت ترکیب پیچیده‌ای از این مسئله است.

فرض کنید ترکیب هر غیرپایانه و پدرش معادل یک تاس n وجهی است. پس به تعداد ترکیب همهٔ غیرپایانه‌ها و پدرشان تاس خواهیم داشت. در مسئلهٔ استخراج گرامر جایگزینی درخت، هر غیرپایانه به بی‌نهایت شکل ممکن می‌تواند گسترش یابد. بنابراین ترکیب هر غیرپایانه و پدر آن نیز به بی‌نهایت شکل ممکن می‌تواند گسترش یابد. درنتیجه، هر تاس ما یک تاس بی‌نهایت وجهی است. درحقیقت هر تاس به این شکل است که با ترکیب یک غیرپایانه و پدرش نام‌گذاری شده و روی هر وجه آن یک درخت اوّلیه ممکن چاپ شده است. با توجّه به این نکته که هر درخت تجزیه می‌تواند با درخت‌های اشتقاق مختلفی تولید شده باشد، می‌توان نتیجه گرفت که داده‌های مشاهده شده می‌تواند ترکیب‌های مختلفی از پرتاب تاس‌ها تولید شده باشد. بنابراین در هر تجزیهٔ نحوی ممکن، از داده‌ها تعداد محدودی از تاس‌های موجود شرکت دارند. حال اگر توزیع احتمال تاس‌هایی را که در تولید داده‌ها شرکت داشته‌اند، بیابیم توزیع احتمال **e** را یافته‌ایم. تصمیم‌گیری راجع به توزیع احتمال هر یک از تاس‌ها یک مسئلهٔ انتخاب مدل است. اگر بخواهیم برای تصمیم‌گیری راجع به توزیع احتمال هر یک از تاس‌ها از مدل بیزی از فرایند شناخت انتخاب مدل استفاده کنیم، باید از تعمیم‌یافتهٔ این مدل برای تاس بی‌نهایت وجهی استفاده کنیم. همان‌طور که در قبل هم اشاره کردیم، این تعمیم همان فرایند دیریکله است. بنابراین حل مسئلهٔ استخراج گرامر جایگزینی درخت، به‌صورت بامربی منجر به تعداد زیادی فرایند دیریکله می‌شود که هر کدام، یک مسئلهٔ انتخاب مدل را برای ما حل می‌کند.

بنابرآنچه در بند قبل توضیح دادیم، به‌ازای ترکیب هر غیرپایانه و پدرش یک فرایند دیریکله خواهیم داشت. اگر فرض کنیم که e' نشان‌دهندهٔ تمامی زیردرخت‌هایی باشد که گرهٔ ریشه و پدر ریشهٔ مشترک دارند، می‌توانیم مدل را به‌صورت زیر بیان کنیم:

$$\mathbf{G_{N,p(N)}}|\boldsymbol{\alpha}_{\mathbf{N,p(N)}}, \mathbf{H_{N,p(N)}}, \mathbf{E} \sim \mathbf{DP}(\boldsymbol{\alpha}_{\mathbf{N,p(N)}}, \mathbf{H_{N,p(N)}}) \quad (۱۲-۱)$$

$$e'|\mathrm{N}, \mathrm{p(N)}, \mathrm{E} \sim \mathrm{G_{N,p(N)}} \quad (۱۲-۲)$$

که در آن N نشان‌دهندهٔ یک غیرپایانه و p(N) نشان‌دهندهٔ پدر آن است. $H_{N,p(N)}$ توزیع پایهٔ فرایند دیریکله و $\alpha_{N,p(N)}$ پارامتر تمرکز این فرایند است. $G_{N,p(N)}$ نیز ن شان‌دهندهٔ متغیر تصادفی از نوع توزیع احتمال است که به ازای ترکیب هر غیرپایانه و پدرش یکی وجود دارد. اگر از دیدگاه مدل استخراج گرامر بخواهیم توصیف کنیم، $G_{N,p(N)}$ تمامی توزیع‌های ممکن برای قوانینی است که دارای ریشهٔ N دارند؛ و پدر ریشهٔ آنها p(N) بوده است. مجموعهٔ $G_{N,p(N)}$ها به‌ازای تمامی N و p(N)های موجود، یک متغیر تصادفی است که دامنهٔ آن تمامی گرامرهای ممکن است.

حال اگر همانند آنچه در رابطه با مدل بیزی از فرایند شناخت انتخاب مدل در حل مسئلهٔ سکه انجام دادیم از پارامتر آزاد مدل انتگرال بگیریم، انتگرالی شبیه رابطهٔ (۱) حاصل خواهد شد. این انتگرال‌گیری طبق فرمول (۶) در این مسئله رابطهٔ زیر را نتیجه خواهد داد:

$$P\left(e'_i \middle| e'_{-i}, N, p(N), \alpha_{N,p(N)}, H_{N,p(N)}, E\right) = \frac{\alpha_{N,p(N)} H\left(e'_i | N, p(N)\right) + cnt_{-i}(e'_i | N, p(N))}{\alpha_{N,p(N)} + cnt_{-i}(e'' | N, p(N))} \quad (۱۳)$$

در رابطهٔ فوق e'_i نمونهٔ زیردرخت iاُمی است که از فرایند دیریکله تولید می‌شود. e'_{-i} بردار $i-1$ نمونهٔ قبلی است که از فرایند دیریکله نمونه‌گیری شده‌اند. $cnt_{-i}(e'_i|N, p(N))$ نیز تابعی است که نشان‌دهندهٔ تعداد رخداد زیردرخت e'_i با گرهٔ ریشهٔ N و پدر ریشهٔ $p(N)$ است که در $i-1$ نمونه‌گیری قبلی تولید شده است. "e" نشان‌دهندهٔ تمامی زیردرخت‌های تولید شده است.

نتیجهٔ انتگرال‌گیری از پارامتر آزاد مدل در فرایند دیریکله، فرایند رستوران چینی را نتیجه خواهد داد. یعنی رابطهٔ فوق رابطهٔ توزیع احتمال انتخاب میز در فرایند رستوران چینی است. به بیان دیگر این توزیع احتمال همان توزیع احتمالی است که پیاده‌ساز حافظه برای ذخیره‌سازی واحدهای پرکاربرد است. اثر برچسب‌زدن هر تاس در مدل بالا با گرهٔ ریشه و پدر ریشهٔ مبتنی‌بر سابقه‌شدن مدل نتیجه‌شده خواهد بود. امّا با این کار تعداد تاس‌های درگیر در مسئله بیشتر خواهد شد. یعنی از دیدگاه فرایند رستوران چینی تعداد رستوران‌ها افزایش پیدا کرده است. درنتیجه این تغییر تعداد میزهایی که در هر رستوران پر می‌شود کم‌تر خواهد بود. این مسئله باعث می‌شود در رابطهٔ بالا مخرج کسر کوچک‌تر و فاصلهٔ بین احتمال‌های کوچک و بزرگ بیش‌تر شده و درنتیجه یادگیری یا ذخیره‌سازی تصادفی بهتر انجام گیرد. یعنی خاصیت "ثروتمند ثروتمندتر می‌شود" بدین ترتیب تشدید می‌شود.

۴- فرایند یادگیری مدل

همان‌طور که در بخش قبل نیز اشاره شد، هدف در مدل ارائه‌شده، یافتن توزیع احتمال $P(e|E)$ است. این توزیع احتمال با توجه به تعریف مدل، به‌صورت زیر تغییر خواهد کرد:

$$P(e|E, \alpha, H)$$

علّت این تغییر آن است که پارامترهای فرایندهای دیریکله ($H_{N,p(N)}$، $\alpha_{N,p(N)}$) متغیرهایی هستند که باید مقادیر آنها در مدل تعیین شود.

α و H برداری از تمامی پارامترهای تمرکز و توزیع اولیه برای تمام فرایندهای دیریکله هستند. این توزیع احتمال توزیع پیچیده‌ای است که راه حل ریاضی برای محاسبهٔ تحلیلی آن نداریم. امّا می‌توانیم با استفاده از راه حل‌های موجود برای تخمین انتگرال‌ها، نظیر مونت کارلوی زنجیرهٔ مارکوف، این توزیع را تخمین بزنیم.

مدل ارائه‌شده، موجب می‌شود توزیع احتمال فوق را بتوان توزیع اشتراک تاس‌ها دانست. یعنی اگر تاس‌ها را متغیرهای تصادفی فرض کنیم، توزیع احتمال فوق توزیع اشتراک این متغیرهای تصادفی است. همان‌طورکه در بخش دو اشاره کردیم، این‌گونه توزیع‌ها را که توزیع احتمال اشتراک تعداد زیادی متغیرهای تصادفی است، با استفاده از روش نمونه‌برداری گیبس که یک نسخه از روش‌های مونت کارلوی زنجیرهٔ مارکوف است، می‌توان تخمین زد. برای این کار باید توزیع یک متغیر را به شرط متغیرهای دیگر محاسبه کنیم و هر بار از یکی از متغیرها به شرط بقیهٔ متغیرها نمونه تولید کنیم تا پس از تولید نمونه برای تمامی متغیرها یک نمونه از توزیع احتمال هدف تولید کرده باشیم *(Resnik, et al., 2010)*. بدین منظور ما از مدل ارائه شده در *(Cohn, et al., 2009)* تبعیت کردیم. بدین شکل که فرض کنید برای هر گرهٔ غیر ریشه در پیکره، یک متغیر باینری در نظر می‌گیریم که مشخص کند این گره محل جایگزینی است یا نه. درحقیقت با درنظرگرفتن تاس‌هایی که گفتیم، مجموعه این متغیرها تعیین می‌کنند که کدام تاس‌ها باید در این مرحله پرتاب شوند.

فرض کنید برای دور اوّل نمونه‌برداری گیبس را به‌صورت تصادفی، گره‌هایی که محل جایگزینی هستند، تعیین کرده‌ایم. یعنی به‌طور تصادفی تاس‌هایی را که در دور اوّل نمونه‌برداری گیبس پرتاب می‌شوند، تعیین کنیم. حال نمونه‌بردار گیبس ما روی تمامی گره‌هایی که در مرحله قبل برچسب جایگزینی نخورده‌اند، حرکت کرده و با تولید نمونه تعیین می‌کند که گرهٔ مورد بررسی، محل جایگزینی بشود یا نه. این طریقه عملکرد سبب می‌شود که عملیات پرتاب هر تاس نیز در دل فرایند نمونه‌برداری گیبس انجام شود. این بدان معناست که فرایند پرتاب یک تاس پس از نمونه برداری از چند گره صورت می‌گیرد. به‌طوری‌که زمانی فرایند پرتاب یک تاس کامل و یک وجه آن ظاهر می‌شود، که در تمامی مسیرها از یک گره جایگزینی به سمت پایین درخت، یک گرهٔ جایگزینی به‌وسیلهٔ نمونه‌برداری تعیین شود. یعنی مرزهای یک درخت اوّلیه تعیین گردد.

حال فرض کنید که می‌خواهیم از متغیر تصادفی مرتبط با گره N نمونه‌گیری کنیم که آیا محل جایگزینی بشود یا نه. همان‌طور که از جملهٔ قبل مشخّص است و قبل از این هم گفتیم این متغیر تصادفی یک متغیر باینری است. برای نمونه‌گیری ابتدا باید توزیع احتمال آن را در این مرحله به شرط بقیهٔ متغیرها تعیین کرده و سپس از آن نمونه‌گیری

کنیم. بسته به مقدار تعیین شده برای این متغیر ممکن است دو حالت اتفاق بیفتد. یکی اینکه گرهٔ مرتبط با این متغیر، محل جایگزینی انتخاب نشود و درنتیجه باید گرهٔ داخلی یک درخت e_m شود. دیگری اینکه این گرهٔ محل جایگزینی تعیین شده و مرز دو درخت e_a و e_b شود *(Cohn, et al., 2009)*. پس باید احتمال رخداد این دو حالت را حساب کرده و یک توزیع دوجمله‌ای[1] به دست آوریم و از آن نمونه تولید کنیم. این توزیع به‌صورت زیر خواهد بود:

$$P\left(e_m \middle| N_{e_m}, p(N_{e_m}), e_{-i}, \alpha_{N_{e_m},p(N_{e_m})}, H_{N_{e_m},p(N_{e_m})}, E\right) \quad (۱۴)$$
$$= \frac{H(e_m|N_{e_m},p(N_{e_m}))\alpha_{N_{e_m},p(N_{e_m})} + cnt_{-i}(e_m|N_{e_m},p(N_{e_m}))}{cnt_{-i}(e"|N_{e_m},p(N_{e_m})) + \alpha_{N_{e_m},p(N_{e_m})}}$$

$$P\left(e_a, e_b \middle| N_{e_a}, p(N_{e_a}), e_{-i}, \alpha_{N_{e_m},p(N_{e_m})}, H_{N_{e_m},p(N_{e_m})}, E\right) \quad (۱۵)$$
$$= \frac{H(e_a|N_{e_a},p(N_{e_a}))\alpha_{N_{e_a},p(N_{e_a})} + cnt_{-i}(e_a|N_{e_a},p(N_{e_a}))}{cnt_{-i}(e"|N_{e_a},p(N_{e_a})) + \alpha_{N_{e_a},p(N_{e_a})}}$$
$$\times \frac{H(e_b|N_{e_b},p(N_{e_b}))\alpha_{N_{e_b},p(N_{e_b})} + cnt_{-i}(e_b|N_{e_b},p(N_{e_b})) + \delta(e_a,e_b)}{cnt_{-i}(e"|N_{e_b},p(N_{e_b})) + \alpha_{N_{e_b},p(N_{e_b})} + \delta((N_{e_b},p(N_{e_b})),(N_{e_a},p(N_{e_a})))}$$

$\mathbf{e_{-i}}$ تمامی درخت‌های اولیه‌ای هستند که درنهایت یک گرهٔ غیرپایانه با $\mathbf{e_a}, \mathbf{e_b}$ یا $\mathbf{e_m}$ اشتراک دارند.

رابطهٔ (۱۴) همان رابطهٔ (۱۳) است که متغیرهای آن با متغیرهای جدید جایگزین شده‌اند. همچنین به‌راحتی قابل مشاهده است که رابطهٔ (۱۵) نیز با استفاده از رابطهٔ (۱۳) و رابطهٔ زیر که قانون زنجیری است، به‌دست می‌آید:

$$P\left(e_a, e_b \middle| N_{e_a}, p(N_{e_a}), e_{-i}, \alpha_{N_{e_m},p(N_{e_m})}, H_{N_{e_m},p(N_{e_m})}, E\right) =$$
$$P\left(e_a \middle| N_{e_a}, p(N_{e_a}), e_{-i}, \alpha_{N_{e_m},p(N_{e_m})}, H_{N_{e_m},p(N_{e_m})}, E\right) \times$$
$$P(e_b|e_a, N_{e_a}, p(N_{e_a}), e_{-i}, \alpha_{N_{e_m},p(N_{e_m})}, H_{N_{e_m},p(N_{e_m})}, E)$$

فرایند استخراج گرامر در ادامه به این شکل خواهد بود که به ازای هر گره، که در گام قبلی به‌عنوان گرهٔ جایگزینی انتخاب شده، روابط بالا محاسبه می‌شود و از آنها نمونه تولید می‌شود. هر نمونه از توزیع‌های فوق یک درخت اوّلیه (قانون گرامر استخراجی) را نتیجه می‌دهد. بدین‌ترتیب پس از هر دور حرکت روی پیکره و تولید نمونه از توزیع‌های فوق، به‌ازای هر گرهٔ جایگزینی، یک تجزیهٔ نحوی از کل جملات پیکره خواهیم داشت. در مرحلهٔ پایانی این فرایند تکراری تجزیهٔ نهایی جملات پیکره تولید خواهد شد. این تجزیهٔ نهایی از درخت‌های اوّلیه‌ای ساخته شده است که درحقیقت قوانین گرامر نهایی هستند.

[1] Binomial

برای محاسبهٔ روابط ارائه شده در این بخش باید پارامترهای فرایندهای دیریکله تعیین شده باشند. ما هر دو پارامتر فرایند دیریکله ($H_{N,p(N)}$، $\alpha_{N,p(N)}$) را به تبعیت از *(Cohn, et al., 2009)* تعیین کردیم. به این ترتیب که H را تابع تصادفی در نظر گرفتیم که روی هر گره، حرکت کرده و یک سکه پرتاب می‌کند و براساس نتیجهٔ سکه تصمیم می‌گیرد که گرهٔ مذکور گسترش پیدا کند یا نه. هر متغیر α نیز از توزیع گاما $Gamma(0.001,1000)$ نمونه‌گیری می‌شود.

فرایند نمونه‌برداری گیبس باید تعداد زیادی نمونه از توزیع احتمال هدف تولید کند تا بتوان با استفاده از این نمونه‌ها تخمین مناسبی از توزیع احتمال هدف تولید کرد. طبق آنچه در *(Goldwater, et al., 2007)* بیان شده، اگر فرایند نمونه‌برداری گیبس را با ذوب فلز شبیه‌سازی شده[2]، همراه کنیم و فرایند کاهش دما را به اندازه کافی کُند انجام دهیم، پس از تولید نمونه‌های زیادی از توزیع احتمال هدف، نمونه‌های تولید شده حول نقطه MAP توزیع احتمال هدف متمرکز می‌شوند. بدین ترتیب نمونهٔ تولید شده را در مرحلهٔ آخر، نمونهٔ MAP به‌حساب‌آورد. این نمونهٔ تولید شدهٔ نهایی همان تجزیهٔ نهایی پیکرهٔ یادگیری است که از قوانین گرامر استخراجی تشکیل شده است.

درصورتی‌که در مدل ارائه‌شده در این تحقیق فرایند دیریکله به فرایند پیتمن-یور تغییر یابد، نسخهٔ مبتنی‌بر سابقهٔ مدل ارائه‌شده در *(Cohn, et al., 2010)* به‌دست خواهد آمد. به همین دلیل ما از پیاده‌سازی انجام شده برای کار *(Cohn, et al., 2010)* استفاده کردیم تا مدل خود را ایجاد کنیم. برای ایجاد مدل خود ما فرایند پیتمن-یور پیاده‌سازی شده را با تغییر پارامترهای فرایند به فرایند دیریکله تبدیل کردیم. سپس با روش برچسب‌گذاری پدر که در (*Johnson, 1998*) ارائه شده است پیاده‌سازی مذکور را به پیاده‌سازی مدل خود تبدیل کردیم.

با توجّه به اینکه الگوریتم یادگیری، در هر تکرار به ازای هر گره تصمیم می‌گیرد که گرهٔ مذکور یک گرهٔ داخلی است یا یک گرهٔ مرزی، هزینهٔ اجرای الگوریتم یادگیری $o(i \times n)$ خواهد بود. در این جا i نشان‌دهندهٔ تعداد تکرار الگوریتم و n نمایان‌گر مجموع تعداد گره‌های درخت‌های تجزیهٔ نحوی موجود در پیکرهٔ آموزشی است.

[2] Simulated Annealing

۵- نتایج و ارزیابی

برای آزمایش مدل ارائه شده باید گرامر حاصل از آموزش مدل را بر روی یک پایگاه درختی، در عملیات تجزیهٔ نحوی جملات به‌کارگرفت. از آنجایی‌که استخراج گرامر، سابقهٔ طولانی در حوزهٔ پردازش زبان طبیعی دارد، منبع استاندارد برای آزمایش تجزیه‌گرهای نحوی در زبان انگلیسی وجود دارد.

علاوه‌بر این، براساس معیارهای استاندارد ارزیابی تجزیهٔ نحوی[1] (*Collins, 1996*) (*Harrison, et al., 1991*) یک ابزار استاندارد به نام EVALB[2] برای ارزیابی تجزیهٔ نحوی جملات ایجاد شده است، که امروزه تمامی تجزیه‌گرهای نحوی از این ابزار برای ارزیابی نتایج خود بهره می‌گیرند. معیارهای استاندارد ارزیابی تجزیهٔ نحوی شامل دقت پرانتزگذاری با برچسب[3]، فراخوانی پرانتزگذاری با برچسب[4]، معیار F و تقاطع پرانتزگذاری[5] است.

منبع استاندارد برای آموزش و آزمایش تجزیه‌گرهای نحوی در زبان انگلیسی پیکرهٔ Penn (*Marcus, et al., 1993*) است. همانند کارهای ارائه شده در (*Charniak, et al., 2005; Charniak, 2000; Cohn, et al., 2009; Cohn, et al., 2010*) بخش WSJ در این پیکره برای آموزش و آزمایش تجزیه‌گرهای نحوی به شکلی که در جدول زیر مشاهده می‌کنید، به‌صورت استاندارد تقسیم‌بندی شده است.

(جدول ۱): تقسیم بندی بخش WSJ از پیکره Penn (Cohn, et al., 2010)

Partition	Sections	Sentences
Training	2-21	33180
Development	22	1700
Testing	23	2416

برای ارزیابی عملکرد مدل، باید آن را روی بخش‌های آموزشی پیکره آموزش داده و روی بخش ارزیابی، آزمایش کرد. برای آزمایش، گرامر استخراج شده توسط مدل، به‌وسیلهٔ یک تجزیه‌گر نحوی به کار گرفته می‌شود تا جملات بخش ارزیابی پیکره، تجزیهٔ نحوی شوند. سپس نتیجهٔ این تجزیهٔ نحوی براساس معیارهای استاندارد ارزیابی تجزیهٔ نحوی (**PARSEVAL**) و با استفاده از ابزار **EVALB** که پیاده‌سازی استاندارد این معیارهاست، مورد ارزیابی قرار می‌گیرد. از آنجایی که ما در نظر داشتیم نتایج خود را با نتایج ارائه شده در (*Cohn, et al., 2010*) مقایسه کنیم، تصمیم گرفتیم از تجزیه‌گرهای ارائه شده در این کار برای تجزیهٔ جملات بخش آزمایشی پیکره استفاده کنیم تا الگوریتم تجزیه‌گر، به‌عنوان یک عامل تغییر در نتایج تجزیه مطرح نشود. تجزیه‌گرهای ارائه شده در (*Cohn, et al., 2010*) ابتدا گرامر جایگزینی درخت استخراج شده را، با استفاده از روش کاهش به گرامر مستقل از متن احتمالاتی[6] (*Goodman, 2002*)، به گرامر مستقل از متن احتمالاتی تبدیل می‌کنند. در این فرایند برخی قوانین که احتمال رخداد کمتری دارند حذف می‌شوند. سپس گرامر مستقل از متن احتمالاتی تولید شده، به‌وسیلهٔ تجزیه‌گر ارائه شده در (*Johnson, et al., 2007 b*) برای تولید تعداد زیادی نمونه از فضای احتمال درخت‌های تجزیه به‌کار گرفته می‌شوند. پس از آن، این نمونه تجزیه‌ها با استفاده از یک پذیرنده که بر اساس الگوریتم متروپلیس-هستینگس[7] (*Cohn, et al., 2009*) و از گرامر جایگزینی درخت ایجاد شده است، قبول یا رد می‌شوند.

۵-۲- آزمایش

برای این آزمایش مدل ارائه شده در این تحقیق و مدل ارائه شده در (*Cohn, et al., 2010*) را روی بخش دو تا ۲۱ بخش WSJ پیکره Penn آموزش دادیم. برای این کار تعداد تکرار فرایند نمونه‌برداری گیبس را ۵۰۰۰ تکرار تنظیم کردیم و الگوریتم شبیه‌سازی ذوب فلز را در ۴۰۰۰ تکرار اوّلیه به کمترین دمای خود رساندیم. در این فرایند دما از بیشینهٔ پنج به مقدار کمینهٔ یک کاهش یافت. سپس گرامر نتیجه شده از مرحلهٔ آموزش را به‌وسیلهٔ تجزیه‌گرهای گزارش شده در (*Cohn, et al., 2010*) برای تجزیهٔ نحوی بخش ۲۲ از WSJ به‌کارگرفتیم. دلیل ما برای انتخاب بخش ۲۲ به جای بخش ۲۳، برای آزمایش، آن بود که بتوانیم نتایج را با نتایج گزارش شده در (*Cohn, et al., 2010*) مقایسه کنیم، چرا که آنها بخش ۲۲ را برای آزمایش بکار گرفته‌اند.

[1] PARSEVAL Measures
[2] http://nlp.cs.nyu.edu/evalb/
[3] Labeled Precision
[4] Labeled Recall
[5] Cross Bracketing
[6] PCFG Reduction
[7] Metropolise-Hastings

(جدول ۲): درصد به دست‌آمده برای معیار F1 برای مدل بیزی شناختی که در این کار معرفی شده و مدل ارائه شده در مقاله *(Cohn, et al., 2010)* برای تجزیه بخش ۲۲ پیکرهٔ Penn. سطر آخر اعداد گزارش شده برای گرامر مستقل از متن احتمالاتی است که با روش مبتنی بر درست‌نمایی بیشینه استخراج شده است.

روش	F1	دقت	فراخوانی	مطابقت کامل
مدل بیزی شناختی [این مقاله]	92.9	92.94	92.99	52.92
کُهن و همکاران *(Cohn, et al., 2010)*	83.8	-	-	26.2
(Cohn, et al., 2010) MLE PCFG	63.1	-	-	6.7

ما مدل خود را با تجزیه‌گر *MPT*[1] *(Cohn, et al., 2010)* مورد آزمایش قرار دادیم. مطابق با آنچه در رابطه با تئوری این تجزیه‌گر در *(Cohn, et al., 2010)* بیان شده است، این تجزیه‌گر با استفاده از روش‌های مونت کارلوی زنجیرهٔ مارکوف، احتمال درخت‌های تجزیهٔ ممکن برای جمله را تخمین می‌زند و محتمل‌ترین درخت تجزیه را برمی‌گزیند. همان‌طور که در ۰ مشاهده می‌شود، مقدار معیار *F1* حاصل از مدل ما، با اختلاف به‌نسبه قابل توجهی بیشتر از مقدار حاصل شده از مدل کُهن و همکاران، با استفاده از تجزیه‌گر *MPT*، است. این موضوع را این‌گونه تحلیل می‌کنیم که با افزایش تعداد رستوران‌ها در مدل ما پیچیدگی توزیع احتمال هر رستوران کاهش پیدا کرده است یعنی تعداد میزهای کمتری در هر رستوران پر شده است. بررسی قوانین گرامر استخراج شده نیز این مسئله را تأیید می‌کند. چرا که قوانینی (درخت‌های اولیه) که دارای ریشهٔ مشترک هستند، در گرامر استخراجی، ما به‌طور متوسط بسیار کم‌تر از قوانین با ریشهٔ مشترک در گرامر استخراجی مدل کُهن و همکاران است. (۰ نشان‌دهندهٔ میانگین تعداد قوانین با ریشهٔ مشترک در گرامرهای استخراجی توسط مدل ما و مدل کُهن و همکاران است.

(جدول ۳): میانگین تعداد قوانین با ریشهٔ مشترک

مدل	میانگین قوانین با ریشه مشترک
مدل بیزی شناختی [این مقاله]	6.35
کُهن و همکاران *(Cohn, et al., 2010)*	24.46

[1] Maximum Probability Tree

کاهش قوانین با ریشهٔ مشترک با توجّه به فرمول (۱۳) موجب اختلاف بیشتر احتمال‌های کوچک و بزرگ برای میزهای (درخت‌های اولیه) هر رستوران می‌شود. همین مسئله، آنطور که درقبل هم گفته شد، موجب تشدید خاصیت "ثروتمند، ثروتمندتر می‌شود" می‌گردد. از دیدگاه شناختی ارائه‌شده در *(O'Donnell, et al., 2009)* ذخیره‌سازی تصادفی در مدل ما نسبت به مدل کُهن و همکاران تقویت شده است و قطعات پر تکرار بهتر تشخیص داده شده و به حافظه سپرده شده‌اند.

به‌منظور اطمینان از عملکرد بهتر مدل ارائه شده و اعتباربخشی بیشتر به ارزیابی مدل، گرامر استخراجی برای تجزیهٔ نحوی بخش استاندارد پیکره Penn (بخش ۲۳) نیز به‌کارگرفته شد. بدین‌ترتیب می‌توان نتایج حاصل از این مدل را با مدل‌های بیشتری نیز مقایسه کرد. البته بسیاری از مدل‌های دیگر استخراج گرامر و تجزیهٔ نحوی سرعت بسیار بالاتری از مدل ارائه‌شده در این مقاله در هنگام تجزیهٔ نحوی جملات فراهم می‌کنند، و سرعت بسیار پایین این مدل در هنگام تجزیهٔ نحوی جملات ضعفی است که حاصل هزینهٔ اجرایی بالای الگوریتم تجزیهٔ نحوی است. هزینهٔ اجرایی این الگوریتم با استفاده از گرامر استخراجی توسط مدل ارائه شده، در اواخر همین بخش محاسبه و گزارش شده است. در (جدول ۴) نتایج آزمایش گرامر حاصل از مدل ارائه‌شده در تجزیهٔ نحوی بخش استاندارد پیکرهٔ Penn (بخش ۲۳) ارائه شده است.

(جدول ۴): درصد به‌دست‌آمده برای معیار F1 برای مدل بیزی شناختی که در این کار معرفی شده و مدل ارائه شده در مقاله *(Cohn, et al., 2010)* برای تجزیه بخش ۲۳ پیکره Penn. سطر آخر اعداد گزارش شده برای گرامر مستقل از متن احتمالاتی است که با روش مبتنی بر درست‌نمایی بیشینه استخراج شده است.

روش	F1	دقت	فراخوانی	مطابقت کامل
مدل بیزی شناختی [این مقاله]	91.82	91.76	91.87	50.75
مدل کُهن و همکاران [آزمایش شده توسط نگارنده]	86.52	86.44	86.59	34.72

(جدول ۵) نیز نتایج حاصل از مدل‌های پیشرو پیشین استخراج گرامر و تجزیه‌ی نحوی را بر روی بخش استاندارد پیکره Penn (بخش ۲۳) گزارش کرده است که می‌تواند با نتایج حاصل از مدل ارائه شده در این مقاله مورد مقایسه قرار گیرد.

(جدول ۵): نتایج حاصل از مدل‌های پیشین استخراج گرامر و تجزیه نحوی روی بخش ۲۳ پیکره Penn.

روش	F1
کالینز *(Collins, 1999)*	88.2
چارنیاک *(Charniak, 2000)*	89.5
چارنیاک و جانسون *(Charniak, et al., 2005)*	91.0
باد *(Bod, 2003)*	90.7
پتروف و کلین *(Petrov, et al., 2007)*	90.0
مک کلاوسکی و همکاران *(McClosky, et al., 2006)*	92.1
مدل بیزی شناختی [این مقاله]	91.82

همان‌طور که مشاهده می‌شود نتیجهٔ حاصل از مدل ارائه‌شده در این مقاله فقط از مدل ارائه شده توسط مک کلاوسکی و همکاران عملکرد ضعیف‌تری داشته است.

از آنجایی‌که فرایند یادگیری و تجزیهٔ نحوی در مدل ارائه شده یک فرایند تصادفی است، لازم است اختلاف مقدار معیار F1 حاصل از مدل ارائه شده و مدل کهن و همکاران [۱۹] از دیدگاه ارزشمندی آماری مورد آزمون قرار گیرند تا بتوان نتیجه گرفت که اختلاف مشاهده شده معنادار است. برای محاسبهٔ ارزشمندی آماری این اختلاف، ما از آزمون t استفاده کردیم. میزان احتمال درستی فرضیهٔ خلف از طریق این آزمون برابر با $1.62 * 10^{-12}$ شد. فرضیهٔ خلف در این آزمون آن بود که مدل ما در تجزیهٔ جملات، بدتر مساوی مدل کهن و همکاران عمل می‌کند. احتمال محاسبه شده برای فرضیهٔ خلف، نشان‌دهندهٔ نادرستی این فرضیه است. بدین ترتیب می‌توان نتیجه گرفت که اختلاف نتایج حاصل از آزمایش مدل ارائه شده و مدل کهن و همکاران *(Cohn, et al., 2010)* از دیدگاه آماری ارزشمند است. لازم به ذکر است که افزایش سطح به کارگیری اطّلاعات سابقه می‌تواند موجب تولید گرامرهای دقیق‌تری شود *(Feili, et al., 2006)* . اما همانطور که در *(Feili, et al., 2006)* اشاره شده است، این کار می‌تواند موجب افزایش‌نمایی در اندازهٔ گرامر تولیدی شود و عملیات تجزیهٔ نحوی را غیر ممکن کند. علاوه‌بر این، افزایش سطح اطلاعات سابقه می‌تواند باعث بروز مشکل تنکی دادهٔ یادگیری نیز بشود و بدین‌ترتیب میزان یادگیری کاهش یابد.

آزمایش‌های انجام شده در این تحقیق، از دیدگاه زمان، آزمایش‌هایی بسیار پرهزینه هستند. دلیل این امر هزینهٔ بالای زمانی الگوریتم تجزیهٔ نحوی MPT، به‌ویژه پس از دخیل‌کردن اطّلاعات سابقه، است. هزینهٔ زمانی اجرای این الگوریتم، به‌صورت زیر بیان می‌شود:

$$O(N^3n^3 + iN^2n^2)$$

که در آن $\mathbf{N}$ تعداد غیرپایانه‌های گرامر، n طول جمله و i تعداد نمونه‌هایست که در فرایند مونت کارلوی زنجیرهٔ مارکوف از درخت تجزیهٔ یک جمله تولید می‌شود. پس از دخیل‌کردن اطّلاعات سابقه در فرایند یادگیری، هزینهٔ تجزیه، به‌صورت زیر تغییر می‌کند:

$$O(N^4n^3 + iN^2n^2)$$

همان‌طور که مشاهده می‌شود، ترتیب زمانی تجزیهٔ نحوی جملات در این حالت $\mathbf{N}$ برابر شده است. در آزمایش‌های انجام‌شده در این تحقیق مقدار $\mathbf{N}$ برابر با ۴۵ است. این موضوع سبب شد در آزمایش‌های انجام شده، زمان تجزیهٔ نحوی جملات پس از دخیل‌کردن اطّلاعات سابقه، به‌شدّت افزایش یابد. بنابراین، برای تجزیهٔ نحوی جملات بخش ۲۲ بانک درختی Penn در زمانی منطقی، مجبور به اجرای موازی الگوریتم روی حدود ۳۸۰ پردازش‌گر موازی شدیم.

۶- نتیجه‌گیری

در این مقاله، ما یک مدل بیزی غیرپارامتری برای استخراج گرامر مبتنی بر سابقه ارائه کردیم. مدل ما علاوه بر مبتنی بر سابقه بودن، تفاوت دیگری نیز با مدل‌های مشابه خود که در *(Cohn, et al., 2010)*، *(Cohn, et al., 2009)* و *(O'Donnell, et al., 2009)* ارائه شده‌اند، دارد. مدل ما مسئلهٔ استخراج گرامر جایگزینی درخت را به تعداد زیادی مسئلهٔ انتخاب مدل تقسیم می‌کند و هر مسئلهٔ انتخاب مدل را با روشی که بر مبنای یک مدل بیزی از فرایند شناختی انتخاب مدل بنا شده است، حل می‌کند. مبتنی بر سابقه‌شدن مدل، موجب ساده‌تر شدن هر مسئلهٔ انتخاب مدل می‌شود و درعوض تعداد این مسائل را در مسئلهٔ اصلی افزایش می‌دهد. ساده‌تر شدن مسائل انتخاب مدل موجب عملکرد بهتر فرایند دیریکله یا فرایند رستوران چینی در فراگیری واحدهای پایهٔ پر کاربرد می‌شود. نتایج حاصل از مدل ما در مقایسه با مدل‌های *(Cohn, et al., 2010)* و *(Cohn, et al., 2009)* در آزمایش‌های مشابه، این موضوع را در عمل نشان می‌دهد. علاوه‌بر این، مدل ما مدل مبتنی بر

انتخاب مدل را نسبت به مدل‌های مبتنی بر درست‌نمایی بیشینه، در حل مسائل انتخاب، مدل نشان می‌دهد.

۷- تشکر و قدردانی

در پایان لازم می‌بینیم از پژوهشگاه دانش‌های بنیادی (IPM) برای فراهم کردن امکان اجرای آزمایش‌های این تحقیق بر روی رایانهٔ خوشه‌ای آن پژوهشگاه (www.cluster.hpc.ipm.ac.ir) تشکر کنیم.

۸- مراجع

Bod Rens, 1992. A computational model of language performance:data oriented parsing. Proceedings COLING'92. - Nantes, France. - pp. 855-859.

Bod Rens, 2003. An efficient implementation of a new DOP model. EACL '03 Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics. - Budapest, Hungary. - Vol. 1. - pp. 19-26.

Bod Rens, 1995. The problem of computing the most probable tree in data-oriented parsing and stochastic tree grammars. Proceeding EACL '95 Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics - .San Francisco, CA, USA.

Brody Samuel, Lapata Mirella, 2009. Bayesian word sense induction. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics - .Stroudsburg, PA, USA . -pp. 103-111.

Cahill Aoife, 2008. Treebank-Based Probabilistic Phrase Structure Parsing. Language and Linguistics Compass. - Vol. 2. - pp. 18-40.

Charniak Eugene, 2000. A maximum-entropy-inspired parser. Proceedings of the First Annual Meeting of the North American chapter of the Association for Computational Linguistics (NAACL 2000). - pp. 132-139.

Charniak Eugene, 1996. Treebank grammars Proceedings of the Thirteenth National Conference on Artificial Intelligence. - Menlo Park. - pp. 1031-1036.

Charniak Eugene, Johnson Mark, 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. - pp. 173-180.

Cohn Trevor, Blunsom Phil and Goldwater Sharon, 2010. Inducing Tree-Substitution Grammars. Journal of Machine Learning Research. - pp. 3053-3096.

Cohn Trevor, Goldwater Sharon and Blunsom Phil, 2009. Inducing Compact but Accurate Tree-Substitution Grammars. NAACL.

Collins Michael, 1996. A new statistical parser based on bigram lexical dependencies. Proceedings of the 34th Meeting of the Association for Computational Linguistics - .Santa Cruz, CA. - pp. 184-191.

Collins Michael, 1999. Head-Driven Statistical Models for Natural Language Parsing. PhD thesis / University of Pennsylvania.

Feili Heshaam and Ghassem-Sani Gholamreza, 2006. Unsupervised Grammar Induction Using History Based Approach. Computer Speech and Language. - Vol. 20. - pp. 644-658.

Goldwater Sharon and Griffiths Thomas L., 2007, A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. The Association for Computational Linguistics. - pp. 744-751.

Goldwater Sharon, Griffiths Thomas L. and Johnson Mark, 2009. A Bayesian framework for word segmentation: Exploring the effects of context. Cognition. - Vol. 112. - pp. 21-54.

Goodman Joshua, 2002. Efficient Parsing of DOP with PCFG-reductions. Data Oriented Parsing / ed. Bod Rens, Sima'an Khalil and Scha Remko. University of Chicago Press.

Griffiths Thomas, Kemp Charles and Tenenbaum Joshua B, 2008. Bayesian models of cognition. The Cambridge Handbook of Computational Psychology / book auth. Sun Ron. Cambridge University Press.

Harrison Philip [et al.], 1991. Evaluating Syntax Performance of Parser/Grammars of English. Natural Language Processing Systems Evaluation Workshop / ed. Neal Jeannette G.and Walter Sharon M. ACL - pp. 71-78.

Johnson Mark, 1998. The Effect of Alternative Tree Representations on Tree Bank Grammars. Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning - .Sydney, Australia. - pp. 39-48.

Johnson Mark, Goldwater Sharo, 2009. Improving nonparametric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, Colorado.

Johnson Mark, Griffiths Thomas L. and Goldwater Sharon, 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. Advances in Neural Information Processing Systems. - Vol 19.

Johnson Mark, Griffiths Thomas L. and Goldwater Sharon, 2007 b. Bayesian Inference for PCFGs via Markov Chain Monte Carlo. Proceedings of the North American Conference on Computational Linguistics (NAACL '07). The Association for Computational Linguistics. - pp. 139-146.

Marcus Mitchell P., Santorini Beatrice and Marcinkiewicz Mary Ann, 1993. Building a large annotated corpus of English: the Penn treebank. Computational Linguistics. - Vol. 19. - pp. 313-330.

McClosky David, Charniak Eugene and Johnson Mark, 2006. Effective Self-Training for Parsing. The Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics (HLT-NAACL 2006). Brooklyn, New York .

Mohri Mehryar, Sproat Richard, 2006. On a Common Fallacy in Computational Linguistics. SKY Journal of Linguistics / ed. Suominen Mickael . - Vol. 19. - pp. 432-439.

O'Donnell Timothy J., Goodman Noah D. and Tenenbaum Joshua B., 2009. Fragment Grammars: Exploring Computation and Reuse in Language. Massachusetts Institute of Technology - .Cambridge: MIT-CSAIL-TR-2009-013.

Petrov Slav and Klein Dan, 2007. Improved Inference for Unlexicalized Parsing. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. - pp. 404-411.

Resnik Philip, Hardisty Eric, 2010. Gibbs Sampling for the Uninitiated. Technical Report. UMIACS.

Teh Yee Whye, 2010. Dirichlet Processes. Encyclopedia of Machine Learning. Springer.

Titov Ivan, Klementiev Alexandre, 2011. A Bayesian Model for Unsupervised Semantic Parsing. The 49th Annual Meeting of the Association for Computational Linguistics (ACL). Portland, USA.

فیلی هشام، ۱۳۸۵. استنتاج استقرایی یک گرامر محاسباتی برای زبان طبیعی با استفاده از روش بی مربی. پایان نامه دکتری / دانشگاه صنعتی شریف. - تهران.

هشام فیلی تحصیلات خود را در مقطع کارشناسی نرم‌افزار در دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شریف با رتبهٔ یک در سال ۱۳۷۶ به پایان رساند. سپس مقاطع کارشناسی ارشد نرم‌افزار و دکتری هوش مصنوعی را به ترتیب در سال‌های ۱۳۷۸ و ۱۳۸۵ در همان دانشکده تکمیل کرد. از سال ۱۳۸۷ تا کنون عضو هیئت علمی دانشکدهٔ مهندسی برق و کامپیوتر دانشکدهٔ فنی دانشگاه تهران است. زمینه‌های تحقیقاتی مورد علاقهٔ ایشان عبارتند از: "پردازش هوشمند متن و زبان طبیعی"، "مترجم ماشینی"، "داده کاوی"، "بازیابی اطّلاعات" و "شبکه‌های اجتماعی".

ایشان به مدّت یک دهه تجربهٔ مدیریت یک تیم نرم‌افزاری در زمینهٔ تولید سامانهٔ مترجم ماشینی را بر عهده دارند. همچنین از ایشان حدود ۵۰ مقاله در نشریات معتبر و کنفرانس‌های بین‌المللی و ملّی به چاپ رسیده است.

نشانی رایانامهٔ ایشان عبارت است از:
hfaili@ut.ac.ir

حمیدرضا قادر تحصیلات خود را در مقطع کارشناسی مهندسی کامپیوتر گرایش نرم‌افزار در دانشکدهٔ مهندسی برق و کامپیوتر دانشگاه تهران در سال ۱۳۸۸ به پایان رساند. سپس در سال ۱۳۹۰ در مقطع کارشناسی ارشد در گرایش هوش مصنوعی از دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت ایران فارغ التحصیل شد. وی هم اکنون عضو تیم تحقیقاتی توسعهٔ مترجم ماشینی در آزمایشگاه پردازش متن و زبان طبیعی دانشگاه تهران است. زمینه‌های تحقیقاتی مورد علاقهٔ وی شامل "پردازش متن و زبان طبیعی"، "مترجم ماشینی"، "یادگیری ماشین" و "ساخت اتوماتیک منابع زبانی" است.

نشانی رایانامهٔ ایشان عبارت است از:
h.ghader@ece.ut.ac.ir

مرتضی آنالویی دانشیار دانشکدهٔ مهندسی کامپیوتر در دانشگاه علم و صنعت ایران است. تحقیقات دانشگاهی و فعّالیت‌های صنعتی ایشان معطوف به فناوری‌های روز در حوزه‌های "شبکه‌های اجتماعی و اقتصادی"، "هوش مصنوعی و علوم شناختی"، " بهینه‌سازی" و "مهندسی زبان‌های طبیعی" می‌باشد.

نشانی رایانامهٔ ایشان عبارت است از:
analoui@iust.ac.ir