دانشکدهٔ مهندسی کامپیوتر

برچسب‌گذاری خودکار اجزای واژگانی کلام

نام دانشجو

مریم امینیان

۸۵۵۲۱۰۷۷

ma_aminian2@yahoo.com

دکتر بهروز مینایی بیدگلی

بهمن ۸۸

چکیده

در این پژوهش‌نامه به بررسی روش‌های مختلف برچسب‌گذاری اجزای واژگانی کلام پرداخته شده است. برچسب‌گذاری اجزای واژگانی کلام، یکی از زیرساخت‌های مهم در بسیاری از کاربردهای پردازش زبان طبیعی مانند ترجمهٔ خودکار، خطایابی املایی و نحوی، تشخیص گفتار و درک متن است.

در این پژوهش‌نامه پس از بررسی اصول روش‌های آماری و روش‌های مبتنی بر گذار، روش برچسب‌گذاری مبتنی بر حافظه مورد بررسی قرار گرفته است. در این روش ترکیبی از رویکردهای آماری و یک الگوریتم یادگیری خودکار با نام یادگیری مبتنی بر حافظه، دیده می‌شود. روش‌های مبتنی بر حافظه در مقایسه با روش‌های آماری و مبتنی بر گذار، از سرعت بالاتری برخوردار بوده، با استفاده از پیکره‌های کوچک نیز به دقت مطلوبی دست می‌یابند. در حالی‌که در دو روش قبلی، برای رسیدن به دقت مناسب باید از پیکره‌های بزرگ بهره گرفت.

واژه‌های کلیدی: برچسب‌گذاری اجزای واژگانی کلام، روش‌های آماری، روش‌های مبتنی بر گذار، برچسب‌گذاری مبتنی بر حافظه، یادگیری مبتنی بر حافظه.

فهرست مطالب

فصل اول

مقدمه‌ای بر برچسب‌گذاری اجزای واژگانی کلام

۱-۱- مقدمه

شاید بتوان اصلی‌ترین هدف هوش مصنوعی را ایجاد یک سامانهٔ جامع برای حل مسأله دانست. سامانه‌ای که قادر باشد رشتهٔ ادراکات خود از محیط را ارزیابی کرده، پس از پردازش و تحلیل آن پاسخ درخور دهد. یکی از اصلی‌ترین کارهایی که هر سامانهٔ هوشمند در تعامل با انسان، باید انجام دهد، پردازش زبان طبیعی[۱] است. پردازش زبان طبیعی در زمینه‌های فراوان و متنوعی کاربرد دارد؛ کابردهایی مانند خطایابی املایی، نحوی و معنایی، بازیابی اطلاعات[۲]، ترجمهٔ خودکار و خلاصه‌سازی خودکار، بخشی از کاربردهایی است که اساس آن‌ها بر پردازش زبان طبیعی است. امروزه حجم منابع و اسناد الکترونیکی مثل مقالات، کتب و اسناد خاص منظورهٔ اداری و دولتی بسیار بیش‌تر از گذشته است. افزایش روزافزون این منابع و سرعت بالای رشد آن‌ها سبب شده است برخی از کاربردهای پردازش زبان طبیعی مثل ترجمه و خلاصه‌سازی خودکار و خطایابی بیش‌تر از پیش مورد توجه قرار گیرند.

---

۱ Natural Language Processing (NLP)
۲ Information Retrieval (IR)

برچسب‌گذاری خودکار اجزای واژگانی کلام[1]، یکی از مراحل زیرساختی در برخی کاربردها مانند بازیابی اطلاعات، تبدیل متن به گفتار و درک متن است. در این مرحله به هر یک از واژه‌ها در متن برچسبی اختصاص می‌یابد که نقش زبانی آن واژه را در آن جملهٔ خاص نشان می‌دهد. از این رو برچسب‌گذاری، یک از مراحل میانی مهم در پردازش زبان طبیعی به شمار می‌آید و تلاش برای افزایش دقت و سرعت روش‌های برچسب‌گذاری از زمینه‌های مورد توجه در بحث پردازش زبان‌ها هستند.

تاکنون تلاش‌های فراوانی در این زمینه انجام شده است و روش‌های برچسب‌گذاری متعددی برای زبان‌های مختلف به کار گرفته شده‌اند. برخی از مهمترین روش‌های برچسب‌گذاری اجزای واژگانی کلام، عبارتند از: مدل پنهان مارکوف [۱]، سامانه‌های بیشینهٔ آنتروپی [۲]، روش برچسب‌گذاری مبتنی بر گذار [۳] و برچسب‌گذاری مبتنی بر حافظه [۴]. در زبان فارسی نیز تاکنون اقداماتی انجام شده است که در بخش انتهایی این فصل به اختصار به آن‌ها اشاره شده، به بررسی مشکلات و تفاوت‌های زبان فارسی با سایر زبان‌ها پرداخته می‌شود.

۱-۲- برچسب‌گذاری اجزای واژگانی کلام

منظور از برچسب‌گذاری اجزای واژگانی کلام، نسبت دادن برچسب‌های واژگانی به واژه‌ها و نشانه‌های به کار رفته در یک متن است. این برچسب‌ها نشان‌دهندهٔ نقش زبانی هر واژه در متن هستند. البته به هر یک از علامت‌های نگارشی موجود در متن نیز رده‌ای از این برچسب‌ها نسبت داده می‌شود که نوع هر یک از آن‌ها را مشخص می‌کند. برچسب‌گذاری اجزای واژگانی کلام یکی از مهم‌ترین مراحل میانی در بسیاری از کاربردهای پردازش زبان طبیعی است؛ به عنوان نمونه می‌توان به ترجمهٔ خودکار، تجزیه و تحلیل نحوی، بازیابی اطلاعات، خطایابی املایی و نحوی و کاربردهای متفاوت متن‌کاوی اشاره کرد. اهمیت برچسب‌گذاری به عنوان یک مرحلهٔ میانی، در این کاربردها از آن جهت است که در بسیاری از موارد میزان دقت برچسب‌های نسبت داده شده، در دقت نهایی تأثیرگذار است.

بسیاری از واژه‌ها از دیدگاه برچسب واژگانی دارای ابهام هستند. به عبارت دیگر در جایگاه‌های مختلف دارای نقش‌های واژگانی متفاوت هستند. منظور از جایگاه، مکان استفادهٔ واژه در جمله و واژه‌های همسایهٔ آن است که در این پژوهش‌نامه از این پس *"بافت"*[2] نامیده می‌شود. برچسب‌گذاری اجزای واژگانی کلام، در واقع عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینهٔ مورد نظر

---

[1] Port Of Speech Tagging (POS Tagging)
[2] Context

است. به عنوان نمونه به نقش واژهٔ "مردم" در دو جملهٔ زیر توجه کنید. این واژه در یک بافت نقش "اسم" و در بافت دیگر نقش "فعل" را پذیرفته است. البته این ابهام به دلیل ظاهر نشدن مصوت‌های کوتاه در فارسی است که به آن ابهام هم‌نگاره‌ها[1] گفته می‌شود.

" مردم با این جریان مخالفت نمودند."

"دیروز از درد دندان مردم."

نقش‌های زبانی، عموماً در دسته‌هایی رده‌بندی می‌شوند. این رده‌بندی ممکن است بسیار دقیق و جزئی باشد و یا کلی و بدون در نظر گرفتن برخی تفاوت‌ها؛ مثلاً می‌توان برچسب افعال را با در نظر گرفتن جزئیاتی مثل شخص، متعدی یا لازم بودن و زمان فعل به دسته‌های دقیق‌تری تقسیم کرد و یا بدون این جزئیات، فقط به نقش زبانی فعل توجه کرد. به رده‌هایی که برای نقش‌های زبانی در یک سامانهٔ برچسب‌گذاری تعریف می‌شود، مجموعهٔ برچسب‌ها[2] گفته می‌شود.

بر این اساس می‌توان تمامی سامانه‌های برچسب‌گذاری موجود را در دو رده دسته‌بندی کرد [۵]:

۱- روش‌هایی که ابتدا به هر واژهٔ موجود در متن، چندین رده از رده‌های موجود در مجموعهٔ برچسب‌ها را نسبت می‌دهند. در این روش‌ها، مرحلهٔ ابهام‌زدایی پس از تخصیص برچسب‌ها انجام می‌شود.

۲- دستهٔ بعدی روش‌هایی هستند که در حین اختصاص برچسب برای هر واژه، عمل ابهام‌زدایی را نیز به صورت هم‌زمان و توأم انجام می‌دهند.

۱-۳- روش‌های برچسب‌گذاری خودکار اجزای واژگانی کلام

برای برچسب‌گذاری اجزای واژگانی کلام روش‌ها و الگوریتم‌های مختلفی وجود دارد. ورودی این الگوریتم‌ها عبارتند از:

۱- مجموعهٔ برچسب‌های در نظر گرفته شده؛ و

۲- رشته‌ای از واژه‌ها که قصد برچسب‌گذاری آن را داریم.

---

[1] Homograph
[2] Tag Set

خروجی این الگوریتم‌ها، بهترین برچسب واژگانی برای هر یک از واژه‌های رشتهٔ ورودی است. منظور از بهترین برچسب، برچسبی‌ست که به نقش واقعی آن واژه در جمله نزدیک‌تر است [۶].

می‌توان روش‌های برچسب‌گذاری اجزای واژگانی کلام را در سه ردهٔ کلّی دسته‌بندی کرد:

۱- روش‌های مبتنی بر قواعد؛

۲- روش‌های آماری محض؛ و

۳- روش‌های مبتنی بر گذار[۱].

۱-۳-۱- پیکرهٔ متنی

تعریفی که برای پیکره[۲] در [۷] ارائه شده، به این صورت است:

"حجم زیادی از داده‌های زبانی که بر اساس معیارهای مشخص برای هدف معینی جمع‌آوری و ذخیره شده‌اند به صورتی‌که نمایندهٔ زبان و یا گویش مورد مطالعه باشند."

به عنوان نمونه تنها پیکرهٔ موجود در زبان فارسی، پیکرهٔ بی‌جن‌خان [۸] است که در آزمایشگاه زبان‌شناسی دانشگاه تهران تهیه شده است. این پیکره، شامل هفت و نیم میلیون واژهٔ برچسب‌خورده است. در ایجاد این پیکره هم از متون رسمی و هم از متون غیررسمی استفاده شده است. این متون برگرفته از اینترنت، پایان‌نامه‌ها، روزنامه‌ها، مجلّات و کتب مختلف می‌باشند. در نسخهٔ پایانی این پیکره، مجموعهٔ برچسب‌های در نظر گرفته شده شامل پانصد و هشتاد و شش برچسب متمایز است که در شانزده ردهٔ اصلی دسته‌بندی می‌شوند [۵].

۱-۳-۲- روش‌های مبتنی بر قواعد

در این روش‌ها، از یک واژه‌نامه[۳] و مجموعهٔ قواعد[۴] زبانی که به وسیلهٔ عامل انسانی تهیه شده‌اند، برای برچسب‌گذاری استفاده می‌شود. واژه‌نامه شامل واژه‌های زبان و مجموعهٔ برچسب‌هایی است که می‌توان به هر واژه نسبت داد. به عبارتی هر یک از مداخل

---

۱ Transformation Based Tagging
۲ Corpus
۳ Lexicon
۴ Rule Set

این واژه‌نامه، نشان‌دهندهٔ برچسب‌هایی است که در بافت‌های مختلف به یک واژه منتسب می‌شوند. منظور از مجموعهٔ قواعد نیز قانون‌هایی است که برای نقش واژه‌ها در یک زبان وجود دارد. به عنوان نمونه، این‌که در زبان انگلیسی پس از صفات ملکی، اسم می‌آید خود یک قاعده در مجموعهٔ قواعد زبان انگلیسی است. روش‌های مبتنی بر قواعد در دو مرحله واژه‌ها را برچسب‌گذاری می‌کنند [۹]:

۱) در مرحلهٔ اول با استفاده از واژه‌نامهٔ موجود، به هر واژه تمامی نقش‌هایی که می‌تواند در جمله داشته باشد نسبت داده می‌شود. به قرار گرفتن یک واژه در چندین رده در مجموعهٔ برچسب‌ها، ابهام[۱] و به چنین واژه‌ای، مبهم[۲] گفته می‌شود؛ و

۲) در مرحلهٔ دوم، مجموعه قواعد روی واژه‌ها اعمال می‌شود تا ابهام واژه‌ها برطرف شود. در بسیاری از موارد با اعمال مجموعهٔ قواعد، می‌توان نزدیکترین برچسب برای نقش واژه را پیدا و ابهام را به طور کامل رفع کرد. ولی در برخی از موارد نیز استفاده از این قواعد تنها تعداد رده‌های برچسب‌ها را کاهش می‌دهد ولی ابهام را به صورت کامل مرتفع نمی‌کند.

این روش یک مشکل مهم دارد و آن این‌که در برخورد با واژگانی که در واژه‌نامه نباشند، قادر به برچسب‌گذاری نخواهد بود . ورود واژه‌های جدید از سایر زبان‌ها و یا ساخت معادل‌ها و تغییر شکل بسیاری از واژه‌های قدیمی‌تر سبب می‌شود دایرهٔ واژگان زبان‌ها همواره در حال رشد و گسترش باشد. پس واژه‌نامهٔ مورد استفاده باید پیوسته روزآمد شود. به همین علّت این روش به تنهایی در برچسب‌گذارها استفاده نمی‌شود بلکه به همراه بسیاری از مدل‌های آماری و الگوریتم‌های یادگیری خودکار در دستهٔ دیگری از روش‌ها که در فصل سه به آن‌ها پرداخته می‌شود، به کار می‌رود.

۱-۳-۳- روش‌های آماری محض

روش‌های آماری از یک پیکرهٔ از پیش برچسب‌خورده استفاده می‌کنند. در این روش‌ها، از توابع آماری و قواعد احتمالی برای پیش‌بینی برچسب یک واژه در متن استفاده می‌شود. اساس بسیاری از روش‌های آماری، بر تخمین احتمال بیشینه[۳] است؛ تخمین احتمال بیشینه، در نظر گرفتن برچسبی برای یک واژه است که احتمال دیده شدن آن، پس از دنباله‌ای خاص از برچسب‌ها بیشینه باشد [۱۰]. منظور از دنباله‌ای خاص از برچسب‌ها، برچسب‌های واژه‌های قبل از یک واژهٔ خاص است. در فصل دوم این پژوهش‌نامه

---

۱ Ambiguity
۲ Ambiguous
۳ Maximum Likelihood Estimation(MLE)

به بررسی جزئیات روش‌های آماری پرداخته می‌شود. به همین علت در این بخش فقط به یک توضیح اجمالی در مورد این روش‌ها بسنده می‌شود.

۱-۳-۴- روش‌های مبتنی بر گذار

این روش‌ها ترکیبی از روش‌های مبتنی بر قواعد و روش‌های آماری هستند. در این طریقهٔ برچسب‌گذاری، ابتدا بر اساس یکی از روش‌های آماری، برچسبی به هر واژه اختصاص داده می‌شود. به همین دلیل در این روش هم به یک پیکره برچسب‌گذاری شده نیاز است. پس از این مرحله مجموعه‌ای از قواعد روی برچسب‌ها اعمال می‌شوند. از این پس به این قواعد، مجموعهٔ قواعد گذار[۱] گفته می‌شود. هر جا که برچسب در نظر گرفته شده در مرحلهٔ اول، با برچسبی که این قواعد برای واژه در نظر می‌گیرند متفاوت باشد، عنصر یادگیری خودکار[۲] وارد عمل می‌شود [۳]. امّا عنصر یادگیری خودکار چه می‌کند؟

در طبقه‌بندی عامل‌ها[۳] در هوش مصنوعی، عامل‌های یادگیرنده[۴] به عامل‌هایی اطلاق می‌شود که قادرند با استفاده از رشتهٔ ادراکاتی که از محیط دریافت می‌کنند قوانین موجود در پایگاه دانش[۵] خود را تغییر داده و یا قانونی به آن‌ها اضافه کنند؛ به عبارتی این عامل‌ها قابلیت سازمان‌دهی مجدد خود را دارند. وظیفهٔ عنصر یادگیری خودکار در این عامل‌ها، یادگیری از مشاهدات برای افزایش معیار کارآیی[۶] است [۱۱]. در روش‌های برچسب‌گذاری خودکار، معیار کارآیی میزان نزدیکی برچسب نسبت داده شده به واژه، با نقش واقعی آن واژه در جمله است. بنابراین در این روش پس از برچسب‌گذاری واژه‌ها (با روش‌های آماری) برچسب‌گذار به سراغ مجموعهٔ قواعد گذار می‌رود و با اعمال قاعدهٔ گذار مناسب، واژه را دوباره برچسب‌گذاری می‌کند. به عبارتی پایگاه دانش سامانهٔ برچسب‌گذاری دوباره سازمان‌دهی شده، عنصر یادگیری خودکار قواعد جدیدی را به مجموعهٔ قواعد گذار می‌افزاید [۳]. در فصل سوم، به بررسی بیشتر روش‌های مبتنی بر گذار پرداخته می‌شود.

۱-۴- برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی

---

۱ Transformation Rules
۲ Machine Learning Element
۳ Agents
۴ Learning Agents
۵ Knowledge Base
۶ Performance Factor

برچسب‌گذاری اجزای واژگانی در زبان فارسی در مقایسه با سایر زبان‌ها مثل زبان انگلیسی تفاوت‌هایی دارد که از تفاوت‌های ماهوی فارسی با سایر زبان‌ها ناشی می‌شود. بعضی از این تفاوت‌ها عبارتند از:

۱- ساخت‌واژهٔ[۱] افعال: افعال در زبان فارسی بر اساس شخص فعل صرف می‌شوند، از این رو علاوه بر بن فعل، دارای وندهای تصریفی هستند. این عامل سبب می‌شود اشخاص مختلف یک فعل، اشکال متفاوتی داشته باشند.

۲- ساخت‌واژهٔ واژه‌های فارسی: در زبان فارسی وندهایی وجود دارند که همگی به واژه‌ها می‌چسبند؛ مثل ضمایر ملکی و نشانه‌های جمع، «یای» نکره و.... یعنی یک واژه با یک نقش می‌تواند به شکل‌های مختلفی در متن ظاهر شود. این در حالی‌ست که در بسیاری از سامانه‌های مبتنی بر روش‌های آماری، این کلمات، متفاوت از یکدیگر در نظر گرفته می‌شوند و همین عامل در دقت روش تأثیرگذار است [۵، ۱۲].

۳- ابهام در ساخت‌واژه: وجود هم‌نگاره‌ها[۲] در زبان فارسی یکی از عوامل ایجاد ابهام است. هم‌نگاره‌ها واژه‌هایی با شکل نوشتاری یکسان اما معانی متفاوت هستند. ظاهر نشدن مصوت‌های کوتاه در متن فارسی، مهم‌ترین دلیل ابهام هم‌نگاره‌هاست. به عنوان مثال دو واژهٔ "مَردُم" و "مُردَم" به علت نیامدن مصوت‌ها دارای شکل نگارشی یکسان اما نقش‌های متفاوت هستند [۱۳].

۴- تشخیص کران واژه: وجود فاصله، عامل اصلی ابهام در زبان فارسی است. دو واژهٔ "می روم" و "می‌روم" دو شکل نوشتاری از یک فعل هستند که در یکی از فاصلهٔ کامل و در دیگری از نیم‌فاصله استفاده شده است، ناتوانی در تشخیص کران واژه، در اینجا سبب ابهام می‌شود [۱۲].

در حوزهٔ برچسب‌گذاری زبان فارسی تا کنون اقدامات زیر صورت گرفته است:

۱- در [۱۴] روشی برای برچسب‌گذاری متون فارسی بر مبنای روش معرفی شده در [۱۵] ارائه شده است. در این روش، ابتدا همسایه‌های یک واژه در دو بردار به نام‌های بردار زمینهٔ چپ و بردار زمینهٔ راست، جمع‌آوری می‌شود. پس از این مرحله، واژه‌ها بر اساس شباهت بردارهای همسایه‌شان، دسته‌بندی می‌شوند. عملیات برچسب‌گذاری، پس از این رده‌بندی انجام می‌شود. در این مرحله، هر طبقه را می‌توان برچسب‌گذاری کرد. تعداد برچسب‌های در نظر گرفته شده در این آزمایش، چهل و پنج برچسب است. دقت بخش خودکار این روش، ۵۷/۵ درصد است.

---

[۱] Morphology
[۲] Homograph

از جمله مشکلات روش فوق می‌توان به این موارد اشاره کرد [۱۴]:

- ناتوانی در ابهام‌زدایی از برچسب‌ها؛
- نداشتن دقت بالا در برچسب‌گذاری صفات و قیود؛
- ناتوانی در برچسب‌گذاری واژه‌هایی با فراوانی کم در متن.

۲- تنها پیکرهٔ متنی موجود در زبان فارسی، پیکرهٔ بی‌جن خان است. در [۱۶] به بررسی دقیق این پیکرهٔ زبانی و مجموعهٔ برچسب‌های در نظر گرفته شده برای آن پرداخته می‌شود. در این روش برخی برچسب‌های کم تکرار حذف و یا در یک گروه دسته‌بندی می‌شوند. سپس با استفاده از روش تخمین احتمال بیشینه، به واژه‌ها برچسب‌هایی اختصاص داده می‌شود. صحت کلی این روش ۹۰/۴۳ درصد اندازه‌گیری شده است.

ابهام هم‌نگاره‌ها یکی از اساسی‌ترین مشکلات برچسب‌گذاری در تمامی زبان‌ها از جمله زبان فارسی است. در زبان فارسی کارهای محدودی در این زمینه انجام شده است. [۱۳] یکی از فعالیت‌های انجام شده در زبان فارسی در زمینهٔ هم‌نگاره‌هاست که تنها به طبقه‌بندی هم‌نگاره‌ها پرداخته است و روشی برای ابهام‌زدایی از آن‌ها معرفی نمی‌کند. البته در این رابطه، در سایر زبان‌ها کارهای فراوانی انجام شده‌است. اقداماتی که یاروسکی[۱] [۱۷] در این زمینه انجام داده است، نمونه‌ای از این موارد هستند.

هستند.

[۱] Yarowsky

فصل دوم

برچسب‌گذاری با روش‌های آماری

۲-۱- مقدمه

یکی از پرکاربردترین مدل‌ها در برچسب‌گذاری اجزای واژگانی کلام، مدل‌های مارکوف هستند. امروزه بسیاری از روش‌های برچسب‌گذاری مثل روش‌های مبتنی بر گذار و روش‌های مبتنی بر حافظه، از روش‌های آماری در مراحل میانی برچسب‌گذاری استفاده می‌کنند و از طریق ترکیب این روش‌ها با سایر الگوریتم‌ها به نتایج مطلوبی دست یافته‌اند [۳، ۴]. اصول ریاضی این روش‌ها، سبب ایجاد پایه‌های نظری قوی برای این دسته از مدل‌های برچسب‌گذاری شده است. از سوی دیگر نتایج استفاده از این مدل‌ها در بسیاری از کاربردها، نتایج مطلوبی است. از این رو روش‌های آماری خصوصاً مدل‌های مارکوف، در گسترهٔ وسیعی از روش‌ها کاربرد دارند [۱].

اساس تمامی مدل‌های مارکوف بر یک ماشین حالت متناهی وزن‌دار[۱] است که زنجیر مارکوف[۲] نامیده می‌شود. از این رو در این فصل، ابتدا به معرفی زنجیر مارکوف پرداخته می‌شود و نیز فرض‌هایی که در مورد مدل‌های مارکوف وجود دارد مطرح می‌شود.

---

[۱] Weighted Finite State Automata
[۲] Markov Chain

ردهٔ دیگری از روش‌های آماری وجود دارند که به آن‌ها روش‌های چندنگاشت[۱] گفته می‌شود. این روش‌ها عموماً با اعمال فرض‌های مارکوف، به گونه‌ای خاص از مدل‌های مارکوف تبدیل می‌شوند. بسیاری از روش‌های پرکاربرد برچسب‌گذاری مثل روش دونگاشت[۲] و روش سه‌نگاشت[۳]، جزء این دسته روش‌ها محسوب می‌شوند. در ادامه فصل پس از معرفی مدل‌های مارکوف، به بررسی این روش‌ها پرداخته می‌شود.

۲-۲- زنجیر مارکوف

یک ماشین حالت متناهی وزن‌دار، توسعه‌ای بر یک ماشین حالت متناهی است که در آن، هر یک از یال‌ها دارای یک احتمال هستند. این عدد نشان‌دهندهٔ احتمال رفتن از مبدأ این یال به مقصد آن است. پس همواره مجموع احتمالاتی که به یال‌های خروجی از یک حالت نسبت داده می‌شود برابر یک خواهد بود. زنجیر مارکوف، یک ماشین حالت متناهی وزن‌دار است که فرض‌های مارکوف در رابطه با آن صادق هستند [۱۸]. اگر فرض کنیم $X = (x_1, x_2 \dots, x_T)$ دنباله‌ای از متغیرهای تصادفی باشد که مقادیر خود را از مجموعهٔ $S = (s_1, s_2 \dots, s_n)$ که همان فضای حالت است دریافت می‌کنند خواص مارکوف که به آن‌ها فرض‌های مارکوف نیز گفته می شود عبارتند از:

۱- ویژگی افق محدود

$$P(x_{t+1} = s_k | x_1, x_2, \dots, x_t) = \mathrm{P}(x_{t+1} = s_k | x_t) \qquad (۱-۲)$$

۲- بی‌حافظگی

$$P(x_{t+1} = s_k | x_t) = P(x_2 = s_k | x_1) \qquad (۲-۲)$$

در این‌صورت دنبالهٔ متغیرهای تصادفی X، یک زنجیر مارکوف خواهد بود. هر زنجیر مارکوف را با یک ماتریس انتقال نشان می‌دهند که درایه‌های آن به صورت رابطهٔ (۲-۳) تعریف می‌شوند [۱۸]:

$$a_{ij} = P(x_{t+1} = s_j | x_t = s_i) \qquad (۳-۲)$$

---

[۱] N-gram Methods
[۲] Bigram
[۳] Trigram

در حقیقت مدل‌های مارکوف، دسته‌ای از مدل‌های تصادفی هستند که یک فرض مشترک در همهٔ آن‌ها وجود دارد؛ طبق این فرض، برای داشتن احتمال دیده شدن یک مدل خاص در آینده نیازی به مقادیر مدل در گذشتهٔ دور نداریم. منظور از مقدار مدل، مقادیر متغیرهای تصادفی است که حالات در ماشین متناهی معرفی شده در قسمت قبل را تشکیل می‌دهند. این جمله اساس انتزاع در همهٔ مدل‌های مارکوف را به سادگی بیان می‌کند. در این مدل‌ها، طبق فرض دوم مارکوف پیش‌بینی حالت بعدی و به دست آوردن تابع احتمال آن به زمان وابسته نیست [۱۰].

۲-۳- استفاده از مدل مارکوف در برچسب‌گذاری

برای حل مسائل به کمک مدل‌های آماری ابتدا باید مسأله در قالب آن مدل آماری بیان شود و اجزای مسأله بر اجزای آن مدل منطبق گردد. در مسألهٔ برچسب‌گذاری می‌توان هر حالت را معادل یکی از برچسب‌های موجود در مجموعهٔ برچسب‌ها در نظر گرفت. در این صورت وزن یال‌ها مشخص‌کنندهٔ احتمال دیده شدن یک برچسب خاص پس از دنباله‌ای از برچسب‌هاست. جدول (۱-۱) نشان‌گذاری مورد استفاده را نمایش می‌دهد. در این جدول منظور از مجموعهٔ آموزش[۱]، بخشی از پیکره برچسب‌گذاری شده است که برای به دست آوردن آماره‌ها از آن استفاده می‌شود.

جدول (۲-۱) نشان‌گذاری در مدل مارکوف [۱۰]

	واژه در موقعیت i
	برچسب واژهٔ i
	واژه‌های رخ داده در موقعیت i تا i+m
	برچسب های $t_i \dots t_{i+m}$ برای واژه‌های $w_i \dots w_{i+m}$
	jاُمین برچسب در مجموعهٔ برچسب‌ها
	jاُمین واژه در مجموعهٔ واژه‌ها

[۱] Training Set

تعداد رخدادهای w^j در مجموعهٔ آموزش	
تعداد رخدادهای t^j در مجموعهٔ آموزش	
تعداد رخدادهای t^k بعد از t^j	
تعداد رخدادهای t^k بعد از t^j و t^j	C
تعداد رخدادهای w^j با برچسب t^k	

پس با استفاده از فرض‌های مارکوف خواهیم داشت:

اگر w_i نشان‌دهندهٔ واژهٔ iاُم در متنی باشد که می‌خواهیم آن را برچسب‌گذاری کنیم و t_i معرف برچسب این واژه در متن باشد می‌توان احتمال شرطی مربوط به فرض افق محدود را به صورت رابطهٔ (۲-۴) بیان کرد:

$$P(t_{i+1}|t_{1,i}) = P(t_{i+1}|t_i) \qquad (۴-۲)$$

این احتمال شرطی بیان می‌کند که احتمال دیده‌شدن یک برچسب خاص پس از دنباله‌ای از برچسب‌ها، تنها به برچسب واژهٔ قبلی وابسته است. با در نظر گرفتن این فرض وابستگی‌هایی با فاصلهٔ زیاد بین برچسب واژه‌ها در نظر گرفته نمی‌شود. به عنوان نمونه در یک جملهٔ فارسی، ارتباط ضمیری که در ابتدای جمله ظاهر می‌شود با فعلی که در انتها آمده نادیده گرفته می‌شود و این فرض درست نیست. ولی با این وجود نتایج حاصل از برچسب‌گذاری با مدل‌های مارکوف قابل قبول است.

شرط بی‌حافظگی نیز بیان می‌کند که احتمال دیده شدن یک برچسب خاص پس از برچسب دیگر به محل قرار گرفتن واژه‌ها در جمله بستگی ندارد؛ به عنوان مثال اگر یک صفت در ابتدای جمله با احتمال ۰/۱ پس از یک اسم ظاهر می‌شود این احتمال در برچسب‌گذاری بقیهٔ جمله یا یک جملهٔ دیگر تغییر نمی‌کند و ثابت فرض می‌شود.

تمام روش‌های برچسب‌گذاری بر مبنای مدل‌های مارکوف، از تابع احتمال رابطهٔ (۲-۵) استفاده می‌کنند:

$$t_i = \text{argmax}_j \, P(t_j|t_{i-1})P(w_i|t_j) \qquad (۵-۲)$$

در تساوی (۲-۵)، احتمال شرطی اول نشان می‌دهد با فرض دانستن برچسب واژهٔ (i-1)اُم، احتمال آمدن یک برچسب خاص برای واژهٔ jاُم چقدر است. به کمک این احتمال، می‌توان احتمال دیده شدن یک دنباله از برچسب‌ها را به دست آورد. عبارت احتمالی دوم نیز نشان می‌دهد اگر ما انتظار دیدن یک برچسب مشخص را داشته باشیم احتمال آمدن یک واژهٔ خاص چقدر است [۱۰].

برای محاسبهٔ احتمال شرطی اول که به آن احتمال انتقال گفته می‌شود، از تساوی (۲-۶) استفاده می‌شود:

$$P(t^k|t^j) = \frac{C(t^j,t^k)}{C(t^j)} \qquad (۶-۲)$$

برای محاسبهٔ احتمال شرطی دوم نیز که معروف به احتمال خروجی است، از تساوی (۲-۷) استفاده می‌شود:

$$P(w^j|t^k) = \frac{C(w^j,t^k)}{C(t^k)} \qquad (۷-۲)$$

تمامی توابعی که تا کنون معرفی شدند احتمال یک برچسب برای یک واژه را محاسبه می‌کنند اما عملاً در تمامی مدل‌های مارکوف، این فرمول گسترش می‌یابد. در واقع هدف، یافتن محتمل‌ترین دنباله از برچسب‌ها برای رشته‌ای از واژه‌ها است که به آن تخمین احتمال بیشینه[۱] گفته می‌شود. این احتمال با استفاده از قاعدهٔ بیز[۲] به صورت تساوی (۲-۸) بیان می‌شود:

$$argmax_{t_{1.n}} P(t_{1,n}|w_{1,n}) = argmax_{t_{1.n}} \frac{P(w_{1,n}|t_{1,n})P(t_{1,n})}{P(w_{1,n})}$$

$$= argmax_{t_{1.n}} P(w_{1,n}|t_{1,n})\, P(t_{1,n}) \qquad (۸-۲)$$

در این مرحله دو فرض دیگر نیز در نظر گرفته می‌شود تا بتوان تساوی (۲-۸) را به صورتی درآورد که با اطلاعات به دست آمده از پیکره برچسب‌خورده قابل محاسبه باشد [۱۰]. این فرض‌ها عبارتند از:

- واژه‌ها مستقل از یکدیگرند.
- هر واژه تنها وابسته به برچسب خودش است.

---

[۱] Maximum Likelihood Estimation(MLE)
[۲] Bayes Law

اگر فرض شود واژهٔ آغاز هر رشته از واژه‌ها مشخص است و داریم $P(t_1|t_0) = ۱/۰$. می‌توان تساوی (۲-۸) را به صورت رابطهٔ (۲-۹) نوشت:

$$\hat{t}_{1,n} = argmax_{t_{1.n}} P(t_{1,n}|w_{1,n}) = \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \qquad (۹-۲)$$

که در آن $\hat{t}_{1,n}$، محتمل‌ترین دنباله از برچسب‌ها برای رشته‌ای از واژه‌ها است. برای محاسبهٔ احتمالات خروجی و انتقال می‌توان از روابط (۲-۶) و (۲-۷) استفاده نمود.

۲-۴- برچسب‌گذارهای دونگاشتی و سه‌نگاشتی

مدل‌های چندنگاشت[۱] دستهٔ پرکاربردی از مدل‌های آماری برای برچسب‌گذاری هستند. در این مدل‌ها برچسب هر واژه به برچسب n واژهٔ قبلی وابسته است. به همین علت این روش‌ها با نام مدل‌های مارکوف درجهٔ (n-۱)شناخته می‌شوند. مدل‌های مارکوف شرح داده شده در بخش ۲-۳ به مدل‌های دونگاشت[۲] مشهور هستند. اما همان‌طور که در قسمت قبلی نیز گفته شد فرض افق محدود مبتنی بر واقعیت نیست. با گسترش وابستگی هر برچسب به برچسب دو واژهٔ قبلی، می‌توان بر دقت مدل و انطباق آن با واقعیت افزود. برچسب‌گذارهایی با این فرض، به برچسب‌گذار سه‌نگاشت و یا مدل‌های مارکوف درجهٔ دو معروف هستند [۱۰].

در این مدل‌ها خاصیت افق محدود به صورت رابطهٔ (۲-۱۰) بیان می‌شود:

$$P(t_{i+1}|t_{1,i}) = P(t_{i+1}|t_i t_{i-1}) \qquad (۱۰-۲)$$

با توجه به رابطهٔ (۲-۱۰) می‌توان تساوی (۲-۱۱) را برای مدل‌های سه‌نگاشت به صورت زیر نوشت:

$$\hat{t}_{1,n} = argmax_{t_{1.n}} P(t_{1,n}|w_{1,n}) = \prod_{i=2}^{n+1} P(w_i|t_i)P(t_i|t_{i-2}t_{i-1}) \qquad (۱۱-۲)$$

و به طریق مشابه برای محاسبهٔ احتمالات خروجی می‌توان از رابطهٔ (۲-۷) استفاده نمود. احتمالات انتقال نیز با اندکی تغییر از رابطهٔ (۲-۱۲) محاسبه می‌شوند:

---

[۱] N-gram
[۲] Bigram

$$P(t^k|t^i t^j) = \frac{C(t^i,t^j,t^k)}{C(t^i,t^j)} \qquad (۱۲-۲)$$

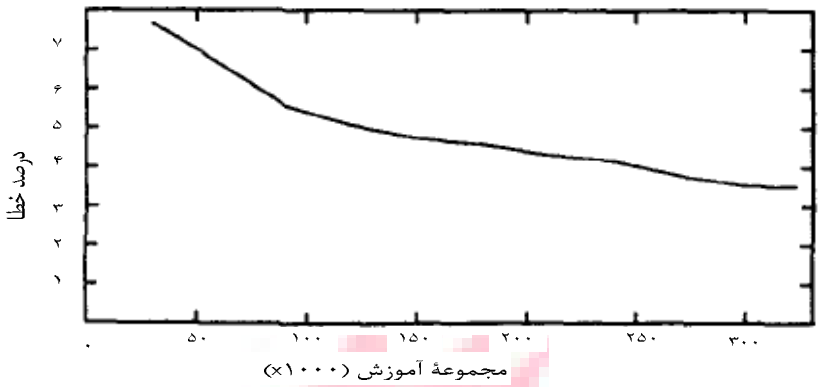
۲-۴- دسته‌بندی مدل‌های مارکوف

مدل‌های مارکوف به دو دسته تقسیم می‌شوند [۱۰]:

- مدل آشکار: در این مدل‌ها، حالات طی شده توسط ماشین حالت متناهی مشخص است. با داشتن یک پیکرهٔ برچسب‌گذاری شده، همواره می‌دانیم ماشین چه حالاتی را برای رسیدن به یک حالت خاص طی کرده است و چه دنباله‌ای از برچسب‌ها به واژه‌های قبلی اختصاص یافته است.
- مدل پنهان: در این دسته از مدل‌های مارکوف، زنجیرهٔ حالات طی شده به صورت قطعی مشخص نیست و حالات مدل به صورت احتمالی تغییر می‌کنند. به همین علت به جای اختصاص یک برچسب برای یک واژه، به دنبال یافتن دنباله‌ای از برچسب‌ها هستیم که احتمال رؤیت آن‌ها بیشینه است. تابع احتمالی که در تساوی (۲-۶) به عنوان احتمال انتقال معرفی شد نشان‌دهندهٔ ماهیت احتمالی گذار بین حالات مختلف است. بنابراین مدلی که در این پژوهش‌نامه معرفی شد، جزء مدل‌های پنهان مارکوف محسوب می‌شود.

۲-۵- نتیجه‌گیری

دقت سامانه‌های برچسب‌گذاری آماری، برای یک مجموعهٔ برچسب ۱۳۰تایی، معمولاً در محدودهٔ ۹۵-۹۷ درصد گزارش شده است [۱۹]. شکل ۲-۱ نیز نشان می‌دهد با افزایش اندازهٔ پیکرهٔ برچسب گذاری شده که به عنوان مجموعهٔ آموزش استفاده می‌شود، درصد خطا در این روش‌ها کاهش می‌یابد.



شکل ۲-۱- منحنی یادگیری برچسب‌گذار آماری [۱۸]

ولی در بررسی روش‌های برچسب‌گذاری دقت یک روش در مقایسه با سایر روش‌هاست که اهمیت می‌یابد. در [۱۹] نتایج برچسب‌گذاری یک روش آماری با نتایج حاصل از یک برچسب‌گذار مبتنی بر قواعد مقایسه شده است. پیکره و مجموعهٔ برچسب‌ها در این آزمایش بین دو روش مشترک است. برچسب‌گذار آماری از مدل مخفی مارکوف و روش سه‌نگاشت برای برچسب‌گذاری استفاده می‌کند، برچسب‌گذار مبتنی بر قواعد نیز سامانهٔ انگ سی‌جی[۱] [۲۰] است. منظور از ابهام در آزمایش که با نسبت تعداد برچسب‌ها بر واژه بیان می‌شود، نشان‌دهندهٔ میانگین تعداد برچسب‌هایی است که به هر واژه منتسب می‌شود. خطای گزارش‌شده برای روش آماری در مقایسه با روش مبتنی بر قواعد در این آزمایش، به ازای تمامی مقادیر این کسر (در فاصلهٔ ۱/۰۰۰ تا ۱/۰۹۳ برچسب/ واژه) بیش‌تر بوده است. از آن‌جایی که مجموعهٔ برچسب‌ها در هر دو روش مشترک است، می‌توان نتیجه گرفت که دقت سامانهٔ مبتنی بر قواعد تا حد زیادی به اطلاعات لغوی و قواعدی که برای رفع ابهام استفاده می‌کنند، وابسته است.

[۱] EngCG

فصل سوم

روش‌های مبتنی بر گذار

۳-۱- مقدمه

روش‌های مبتنی بر گذار که به دلیل فعالیت‌های علمی اریک بریل[۱] در این زمینه، به روش‌های برچسب‌گذاری بریل[۲] معروف هستند، دستهٔ دیگری از روش‌های برچسب‌گذاری اجزای واژگانی کلام می‌باشند. در این روش‌ها با استفاده از برخی رویکردهای آماری، واژگان موجود در پیکرهٔ متنی برچسب‌گذاری می‌شود. سپس از مجموعه‌ای از قواعد برای برچسب‌گذاری دوبارهٔ واژگان استفاده می‌شود؛ به نحوی که میزان خطا در تخصیص برچسب‌ها، دارای سطح دقتی[۳] باشد که از پیش برای سامانهٔ برچسب‌گذاری در نظر گرفته شده است. در همهٔ این روش‌ها از یادگیری مبتنی بر گذار[۴] استفاده می‌شود [۳]. ابتدا به بررسی یادگیری مبتنی بر گذار به عنوان اساس تمامی این روش‌ها پرداخته می‌شود.

۳-۲- یادگیری مبتنی بر گذار

---

۱ Eric Brill
۲ Brill Tagging Method
۳ Significance Level
۴ Transformation Based Learning

همان‌طور که در فصل اول نیز اشاره شد، عامل‌های یادگیرنده در هوش‌مصنوعی، عامل‌هایی هستند که می‌توانند رشتهٔ ادراکات را از محیط دریافت کرده، بر اساس آن‌ها قوانین موجود در پایگاه دانش[۱] خود را تغییر دهند. این عامل‌ها قابلیت سازمان‌دهی مجدد را برای قوانین موجود در پایگاه دانش و یا افزودن قانونی را به این مجموعه دارند. به عبارت دیگر این عامل‌ها می‌توانند از محیط یاد بگیرند. الگوریتم‌های یادگیری بر اساس نوع تعاملی که عامل یادگیرنده با محیط دارد، دسته‌بندی می‌شوند. یادگیری باناظر[۲]، دسته‌ای از الگوریتم‌های یادگیری خودکار هستند که در آن‌ها ردهٔ مجموعه‌ای از نمونه‌ها (مجموعهٔ آموزش) داده شده است و رده‌بندی سایر نمونه‌ها با یادگیری از مجموعهٔ آموزش انجام می‌شود. هدف اصلی در این الگوریتم‌ها، یافتن ردهٔ مناسبی برای یک نمونهٔ جدید است. به همین دلیل این روش‌ها به روش‌های رده‌بندی[۳] معروف هستند [۱۱]. به طریق مشابه، در روش‌های مبتنی بر گذار، یک پیکرهٔ برچسب‌گذاری شده که همان مجموعهٔ آموزش باشد، موجود است، که برچسب واژه‌ها درآن مشخص شده است. برچسب‌گذاری واژه‌ها و جملات ورودی، با در دست داشتن این مجموعهٔ آموزش و یادگیری قوانین گذار از آن، انجام می‌شود. به عبارتی برچسب‌گذاری واژه‌ها با یادگیری از پیکرهٔ آموزشی انجام می‌شود. از این رو یادگیری مبتنی بر گذار، نوعی یادگیری باناظر[۴] محسوب می‌شود. معمولاً برای ارزیابی دقت روش، از یک مجموعهٔ آزمون[۵] استفاده می‌شود.

در روش‌های برچسب‌گذاری مبتنی بر گذار، یک پیکرهٔ متنی برچسب‌گذاری شده موجود است. از آن‌جایی که برچسب واژها در این پیکره به صورت دستی و با استفاده از عامل انسانی تخصیص داده شده است، این برچسب‌ها معیار صحت در این الگوریتم به حساب می‌آیند [۳]. از این پیکرهٔ برچسب‌گذاری شده به عنوان مجموعهٔ آموزش برای یادگیری قوانین استفاده می‌شود. اما معمول آن است که برای برچسب‌گذاری یک پیکرهٔ متنی بخشی از آن را به عنوان مجموعهٔ آموزش استفاده نموده و از قوانین گذار استخراج شده، برای برچسب‌گذاری بقیهٔ متن استفاده می‌کنند.

همان‌طور که در شکل ۳-۱ مشاهده می‌شود، در روش یادگیری مبتنی بر گذار، ابتدا متن بدون برچسب با استفاده از یکی از روش‌های آماری برچسب‌گذاری می‌شود. در [۳]، از برچسب با بیش‌ترین فراوانی[۶] در میان برچسب‌های مختلف یک واژه،

---

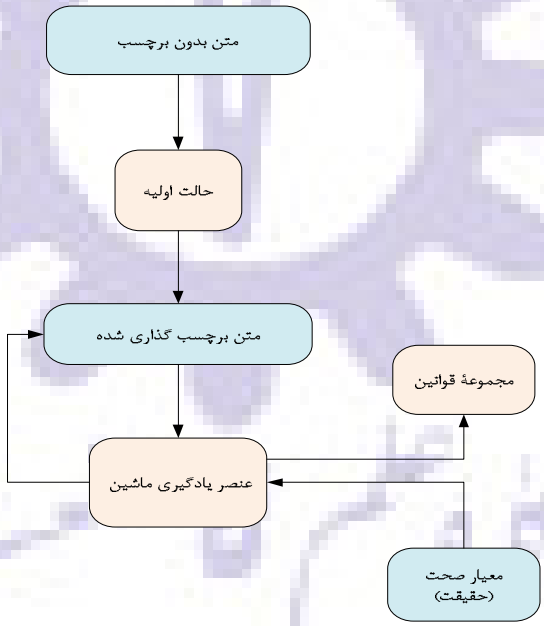
۱ Knowledge Base
۲ Supervised Learning
۳ Classification
۴ Supervised Learning
۵ Test Set
۶ Most Likelihood

برای برچسب‌گذاری استفاده می‌شود. پس از عبور از مرحلهٔ اولیه، متن برچسب‌گذاری شده با نسخهٔ دیگری از همین متن که از قبل به صورت دستی برچسب‌گذاری شده است، مقایسه می‌شود. این متن برچسب‌گذاری شده، همان معیار صحت در این شکل است. خروجی این مرحله، دنبالهٔ مرتب شده‌ای از قوانین گذار است که با اعمال آن‌ها به متن برچسب‌گذاری شده، می‌توان برچسب واژه‌ها را هر چه بیش‌تر به برچسب صحیح، نزدیک کرد.

هر قانون گذار شامل دو جزء اصلی می‌شود [۳]:

۱- قاعدهٔ بازنویسی[۱]: مشخص‌کنندهٔ برچسبی‌ست که باید با برچسب نادرست جایگزین شود.

۲- محیط اعمال[۲]: شروطی را بیان می‌کند که باید در واژه‌های اطراف و یا برچسب آن‌ها وجود داشته باشد تا بتوان قاعدهٔ بازنویسی را اعمال نمود.



شکل ۳-۱- یادگیری مبتنی بر گذار [۳]

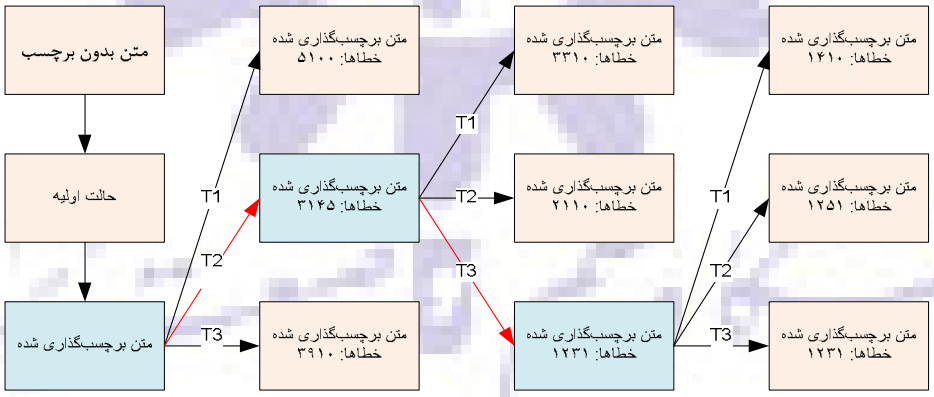
با مقایسهٔ برچسب هر یک از واژه‌ها با برچسب درست آن‌ها در متن برچسب‌گذاری شده، عنصر یادگیری خودکار یک قانون گذار به مجموعهٔ قوانین[۱] اضافه می‌کند. سپس از یک الگوریتم جستجوی حریصانه برای یافتن فهرست مرتب شدهٔ قوانین گذار

[۱] Rewrite Rule
[۲] Triggering Environment

استفاده می‌شود؛ در هر یک از تکرارهای این الگوریتم از میان همهٔ قوانین گذار، قانونی انتخاب می‌شود که اعمال آن بر متن، کمترین تعداد خطا را به دنبال داشته باشد. سپس این قانون به کل متن اعمال می‌شود. این تکرارها تا جایی ادامه پیدا می‌کند که اعمال قوانین گذار بر متن، به کاهش تعداد خطاها منجر نشود، یا این‌که تعداد خطاها، کمتر مساویِ سطح دقت مورد نظر برای برچسب‌گذاری شود [۳].

شکل ۳-۲ یک نمونهٔ ساده از کاربرد این الگوریتم را با این فرض که در مجموعهٔ قوانین تنها سه قانون گذار موجود باشد، نشان می‌دهد. تابع هدف[۲] در هر الگوریتم یادگیری با ناظر، تابعی است که میزان نزدیکی یک حالت را به معیار صحت (حقیقت) نشان می‌دهد. در این مثال تابع هدف مشخص کنندهٔ تعداد خطاهایی است که در برچسب‌گذاری متن با استفاده از یک قانون گذار خاص حاصل می‌شود. از این رو به این روش یادگیری، روش مبتنی بر گذار با هدایت خطا[۳] نیز گفته می‌شود [۳].

همان‌طور که در شکل ۳-۲ مشاهده می‌شود، بعد از اعمال قوانین گذارِ شمارهٔ دو و سه بر متن، با اعمال هیچ یک از قوانین تعداد خطاها کاهش نمی‌یابد. به همین دلیل الگوریتم در این مرحله متوقف می‌شود.



شکل ۳-۲- یادگیری مبتنی بر گذار با هدایت خطا [۳]

---

۱ Rule Set
۲ Objective Function
۳ Transformation-Based Error-Driven Learning

پس می‌توان برای برچسب‌گذاری پیکره به این ترتیب عمل نمود: ابتدا هر واژه با محتمل‌ترین برچسب برای آن واژه، برچسب‌گذاری می‌شود. سپس قانون گذار شمارهٔ دو و پس از آن قانون سه بر کل متن اعمال می‌گردد. با اعمال این قوانین تعداد خطاها در برچسب‌گذاری اولیه کمینه می‌شود. در هر یک از تکرارهای الگوریتم یادگیری، ابتدا تمام محیط‌های اعمال برای یک قانون گذار خاص شناسایی می‌شود. سپس تعداد خطاها با فرض اعمال آن قانون به دست می‌آید. با تکرار این مرحله برای هر یک از قوانین گذار، می‌توان قانونی را که به کمینه شدن تعداد خطاها می‌انجامد انتخاب نموده، آن را به دنبالهٔ مرتب شدهٔ قوانین گذار اضافه کرد.

برای تشخیص نحوهٔ عملکرد هر سامانهٔ برچسب‌گذاری که از این روش استفاده می‌کند، بایستی موارد زیر را مشخص نمود [۳]:

- روشی که برای برچسب‌گذاری متن در حالت اولیه مورد استفاده قرار می‌گیرد؛
- قوانین گذاری که استفاده از آن‌ها در تکرارهای این الگوریتم یادگیری مجاز است؛ و
- تابع هدفی که برای ارزیابی هر حالت و مقایسه با معیار صحت استفاده می‌شود.

در مواردی که اعمال یک قانون گذار بر متن، بر نتیجهٔ اعمال همان قانون در تکرارهای بعدی الگوریتم تاثیرگذار باشد، بایستی این نکته را نیز مشخص نمود که اعمال قوانین بر متن، بلافاصله پس از انتخاب هر قانون انجام می‌شود و یا اینکه دنبالهٔ قوانین در انتها پس از بررسی تمامی قوانین، بر متن اعمال می‌شوند. در روش معرفی شده در [۳]، هر قانون پس از انتخاب بر کل متن اعمال می‌شود.

۳-۳- برچسب‌گذاری مبتنی بر گذار

قوانین گذاری که با استفاده از الگوریتم یادگیری مبتنی بر گذار به دست می‌آیند دارای انواع مختلفی هستند [۳]:

- قوانین گذار مستقل از واژگان[۱]
- قوانین گذار وابسته به واژگان[۱]

[۱] Nonlexicalized Transformation Rules

۳-۳-۱- قوانین گذار مستقل از واژگان

در محیط اعمال این نوع قوانین گذار، به واژه‌ها در متن هیچ ارجاعی داده نمی‌شود و تنها برچسب واژه‌های اطراف یک واژه هستند که در محیط اعمال قوانین گذار مورد توجه قرار می‌گیرند. آن‌چه در این بخش و بخش بعدی مطرح می‌شود الگوی برخی از این قوانین است، که الگوریتم یادگیری از مجموعهٔ آموزش استخراج می‌کند. نمونه‌ای از قوانین گذار مستقل از واژگان به صورت زیر است [۳]:

برچسب (الف) را به برچسب (ب) تغییر بده، اگر:

۱- برچسب واژهٔ قبلی و یا واژهٔ بعدی، برچسب (ج) باشد؛

۲- برچسب واژه‌ای که دو واژه با این واژه فاصله دارد (قبل و یا بعد)، برچسب (ج) باشد؛

۳- برچسب یکی از دو واژهٔ قبلی و یا بعدی، برچسب (ج) باشد؛

۴- برچسب یکی از سه واژهٔ قبلی و یا بعدی، برچسب (ج) باشد؛

۵- برچسب واژهٔ قبلی (ج) و برچسب واژهٔ بعدی (ه) باشد؛ و

۶- برچسب واژهٔ قبلی (ج) و برچسب واژه‌ای که دو واژه با این واژه فاصله دارد (قبل و یا بعد)، برچسب (ه) باشد.

در این الگو برای قوانین گذار، (الف)، (ب)، (ج) و (ه) برچسب‌های مجاز هستند.

۳-۳-۲- قوانین گذار وابسته به واژگان

در قوانین گذار مستقل از واژگان، برچسب هر واژه تنها وابسته به برچسب واژه‌های اطراف فرض می‌شود. این فرض سبب می‌شود ارتباط بین برچسب یک واژه با سایر واژه‌ها در نظر گرفته نشود. در قوانین گذار وابسته به واژگان در محیط اعمال یک قانون، علاوه بر برچسب سایر واژه‌ها، به خود واژه‌های اطراف نیز توجه می‌شود. نمونه‌ای از این قوانین گذار به صورت زیر است [۳]:

برچسب (الف) را به برچسب (ب) تغییر بده، اگر:

---

[۱] lexicalized Transformation Rules

۱- واژهٔ قبلی و یا بعدی، واژهٔ (ج) باشد؛

۲- واژه‌ای که دو واژه با این واژه فاصله دارد (قبل و یا بعد)، واژهٔ (ج) باشد؛

۳- یکی از دو واژهٔ قبلی، واژهٔ (ج) باشد؛

۴- خود واژهٔ مورد نظر واژهٔ (ج)، و واژهٔ قبلی (ه) باشد؛

۵- خود واژهٔ مورد نظر واژهٔ (ج) و برچسب واژهٔ قبلی (و) باشد؛ و

۶- واژهٔ قبلی واژهٔ (ج) و برچسب آن (و) باشد.

در این الگو نیز، منظور از (الف)، (ب) و (و) برچسب‌های مجاز و (ج) و (ه) واژه‌های موجود در متن است.

نتایج آزمایش در [۳] نشان می‌دهد که استفاده از قوانین وابسته به واژگان در مقایسه با قوانین مستقل از واژگان، سبب افزایش سطح دقت برچسب‌گذاری و کاهش میزان خطاها می‌شود.

۳-۴- برچسب‌گذاری واژه‌های ناشناخته

در تمامی روش‌های برچسب‌گذاری، که به نحوی از یک پیکرهٔ از پیش برچسب‌گذاری شده استفاده می‌کنند، مشکل مواجهه با واژه‌های ناشناخته وجود دارد. اگر واژه‌ای قبلاً در پیکره ظاهر نشده باشد، نمی توان از پیکرهٔ آموزش اطلاعات دقیقی راجع به آن واژه به دست آورد. حتی نمی‌توان از توزیع احتمال واژه‌های موجود در پیکره استفاده کرد زیرا توزیع واژه‌های ناشناخته کاملاً متفاوت است.

یک روش ساده برای شناسایی توزیع واژه‌های ناشناخته، در [۲۱] ارائه شده است. ایدهٔ اصلی این روش آن است که احتمال یک برچسب خاص برای یک واژهٔ ناشناخته، با استفاده از توزیع احتمال واژه‌هایی که تنها یک بار در پیکره ظاهر شده‌اند تخمین زده شود. در حقیقت توزیع واژه‌های ناشناخته بسیار شبیه واژه‌هایی است که تنها یک بار در متن ظاهر شده‌اند.

در برچسب‌گذاری مبتنی بر گذار نیز برای برچسب‌گذاری واژه‌های ناشناخته روش‌های مختلفی وجود دارد. روشی که در ادامه به آن اشاره می‌شود، برچسب‌گذاری واژه‌های ناشناخته در یک سامانهٔ برچسب‌گذاری مبتنی بر گذار برای زبان انگلیسی را نشان می‌دهد [۳]:

در اولین مرحله برای برچسب‌گذاری واژه های ناشناخته، اگر واژه با حروف بزرگ الفبای انگلیسی آغاز شوند برچسب "اسم مفرد خاص" و در غیر این‌صورت برچسب "اسم مفرد عام" به آن‌ها اختصاص داده می‌شود. در تکرارهای الگوریتم یادگیری نیز از مجموعهٔ قوانین گذار زیر برای تصحیح برچسب‌ها استفاده می‌شود:

برچسب واژهٔ ناشناخته را از (الف) به (ب) تغییر بده اگر:

۱- با حذف پیش‌وند و یا پس‌وندی به طول حداکثر چهار نویسه، حاصل یک واژهٔ شناخته شده باشد؛

۲- نویسه‌های آغازین واژه، نویسه‌های خاصی باشند؛

۳- با افزودن پیش‌وند و یا پس‌وندی به طول حداکثر چهار نویسه، حاصل یک واژهٔ شناخته شده باشد؛

۴- نویسه‌ای خاص در این واژه ظاهر شده باشد؛ و

۵- با افزودن واژه‌ای خاص به سمت چپ این واژهٔ ناشناخته، حاصل یک واژهٔ شناخته شده باشد.

همان‌طور که مشاهده می‌شود محیط اعمال این قوانین گذار، به واژه‌ها و یا برچسب آن‌ها وابسته نیست. در این الگو، قوانین گذار تنها به ساخت‌واژه‌ها[۱] مرتبط هستند. نکته‌ای که بایستی مورد توجه قرار گیرد آن است که در این‌جا منظور از وند، الزاماً وندهای زبان‌شناسی نیست و وندها می‌توانند هر دنباله‌ای از نویسه‌ها در زبان باشند.

۳-۵- بررسی نتایج برچسب‌گذاری با روش‌های مبتنی بر گذار

نتایج آزمایشات انجام شده در [۳] نشان می‌دهد اگر از یک الگوی قوانین گذار، برای برچسب‌گذاری واژه‌های شناخته شده و ناشناخته استفاده شود، دقت کلی برچسب‌گذاری کمتر از حالتی خواهد بود که برچسب‌گذاری واژه‌های ناشناخته با قوانین گذار خاص خود انجام پذیرد. بنابراین برای افزایش دقت کلی، برای واژه‌های شناخته شده از قوانینی استفاده می‌شود که دقت همین بخش را افزایش دهند. به همین ترتیب برای واژه‌های ناشناخته از قوانین متمایزی استفاده می‌شود که افزایش‌دهندهٔ دقت در این بخش به صورت مجزا باشند. در این آزمایش، از یک پیکره با ۱/۱ میلیون واژه استفاده شده است. از این واژه‌ها، ۹۵۰ هزار واژه در مجموعهٔ آموزش و ۱۵۰ هزار واژه در مجموعهٔ آزمون قرار دارند. از ۶۰۰ هزار واژهٔ موجود در مجموعهٔ آموزش برای یادگیری قوانین گذار مربوط به واژه‌های شناخته شده و از ۳۵۰ هزار واژهٔ باقیمانده برای یادگیری قوانین برچسب‌گذاری

---

[۱] Morphology

واژه‌های ناشناخته استفاده می‌شود. پس از برچسب‌گذاری مجموعهٔ آزمون، دقت برچسب‌گذاری واژه های ناشناخته ۸۲/۲ درصد، و دقت کلی برچسب‌گذاری ۹۶/۶ درصد بوده است.

فصل چهارم

برچسب‌گذاری مبتنی بر حافظه

۴-۱- مقدمه

برچسب‌گذاری مبتنی بر حافظه، یکی از روش‌های برچسب‌گذاری است که از یادگیری مبتنی بر حافظه[۱] استفاده می‌کند. یادگیری مبتنی بر حافظه، یکی از روش‌های یادگیری خودکار است که به صورت باناظر عمل نموده، از استدلال مبتنی بر تشابه[۲] استفاده می‌کند. در این روش، مجموعه‌ای از نمونه‌ها[۳] که همان مجموعهٔ آموزش باشد در حافظه نگه‌داری می‌شود. هر نمونه در حافظه، به صورت یک بردار از ویژگی‌ها[۴] نمایش داده می‌شود. یکی از ویژگی‌های موجود در این بردار، ردهٔ مربوط به نمونه است. برای رده‌بندی یک نمونهٔ جدید، فاصلهٔ آن با نمونه‌های موجود در حافظه محاسبه شده، در نهایت رده‌ای برای نمونهٔ آزمایشی در نظر گرفته می‌شود که حداقل فاصله را با نمونه یا نمونه‌های موجود در حافظه داشته باشد. به عبارتی در این روش، رده‌بندی نمونهٔ جدید بر اساس نمونه‌های موجود در حافظه و ردهٔ آن‌ها انجام می‌شود. از این رو یادگیری مبتنی بر حافظه، نوعی یادگیری استنباطی[۵] باناظر است [۴].

۴-۲- یادگیری مبتنی بر حافظه

---

۱ Memory Based Learning
۲ Similarity based Reasoning
۳ Cases
۴ Vector Of Features
۵ Inductive

در تمامی روش‌های یادگیری با ناظر از جمله روش یادگیری مبتنی بر حافظه، به یک پیکرهٔ برچسب‌گذاری شده، به عنوان مجموعهٔ آموزش نیاز است. در برچسب‌گذاری با این رویکرد، مجموعه‌ای از نمونه‌ها در حافظه نگه‌داری می‌شوند. هر نمونه شامل واژه، اطلاعاتی راجع به بافت چپ[۱] و بافت راست[۲] آن واژه و رده‌ای‌ست که در مجموعهٔ آموزش برای آن در نظر گرفته شده است. این اطلاعات همان بردار ویژگی‌هاست که برای هر نمونه جداگانه نگه‌داری می‌شود [۴].

برای تعیین ردهٔ یک نمونهٔ جدید، این نمونه با نمونه‌های موجود در حافظه مقایسه شده و ردهٔ نزدیک‌ترین نمونهٔ آموزشی به عنوان ردهٔ آن پیش‌بینی می‌شود.

۴-۲-۱- معیار شباهت

برای محاسبهٔ فاصلهٔ یک نمونهٔ جدید با نمونه‌های آموزشی، به یک معیار شباهت[۳] یا معیار فاصله[۴] نیاز است. دقت برچسب‌گذاری این روش و کارآیی آن وابستگی زیادی به نحوهٔ انتخاب این معیار شباهت دارد. رابطهٔ (۴-۱) یکی از ساده‌ترین روابطی‌ست که به عنوان معیار فاصله بین نمونه‌ها، مورد استفاده قرار می‌گیرد . در رابطهٔ (۴-۱)، X و Y دو نمونه‌ای هستند که با هم مقایسه می‌شوند. n تعداد ویژگی‌های موجود در بردار هر یک از نمونه‌ها، و $\sigma(x_i, y_i)$ فاصلهٔ بین ویژگی iاُم دو نمونه است که طبق رابطهٔ (۴-۲) محاسبه می‌شود. از این تابع برای محاسبهٔ فاصله بین ویژگی‌های غیر عددی استفاده می‌شود.

$$\Delta(X,Y) = \sum_{i=1}^{n} \sigma(x_i, y_i) \qquad (۱-۴)$$

$$\sigma(x_i, y_i) = \begin{cases} 1 & \text{اگر } x_i = y_i \\ 0 & \text{در سایر موارد} \end{cases} \qquad (۲-۴)$$

در معیار شباهتی که رابطهٔ (۴-۱) معرفی می‌کند، تمام ویژگی‌های یک نمونه ارزش یکسانی در تعیین ردهٔ نمونه‌های جدید دارند. این درحالی‌ست که برخی ویژگی‌ها، در رده‌بندی از وزن بیشتری برخوردار هستند؛ به عنوان مثال در تعیین ردهٔ یک

---

[۱] Left Context
[۲] Right Context
[۳] Similarity Metric
[۴] Distance Metric

واژه، خود واژه در مقایسه با سایر ویژگی‌های مربوط به بافت، اهمیت بیشتری دارد. از این‌رو هر ویژگی با بهرهٔ اطلاعاتی[۱] آن ویژگی، وزن‌دهی می‌شود [۴].

بهرهٔ اطلاعاتی هر ویژگی نشان می‌دهد مقدار آنتروپی اطلاعات حاصل از مجموعهٔ آموزش، در صورت دانستن مقدار آن ویژگی، به صورت میانگین چقدر کاهش می‌یابد [۴]. با مقایسهٔ بهرهٔ اطلاعاتی ویژگی‌های مختلف، می‌توان تصمیم گرفت کدام ویژگی اطلاعات بیشتری را برای رده‌بندی فراهم نموده و سریع‌تر ردهٔ نمونه را متمایز می‌کند. معیار شباهت وزن‌دار به صورت رابطهٔ (۴-۳) بیان می‌شود که در آن $G(f_i)$ بهرهٔ اطلاعاتی ویژگی iاُم را مشخص می‌کند.

$$\Delta(X,Y) = \sum_{i=1}^{n} G(f_i)\ \sigma(x_i, y_i) \qquad (۳-۴)$$

۴-۳- استفاده از درخت تصمیم برای نگه داری نمونه‌ها در حافظه

یادگیری مبتنی بر حافظه برای برچسب‌گذاریِ نمونه‌های آزمایشی، دارای پیچیدگی زمانی و مکانی بالایی است؛ زیرا برای رده‌بندی هر نمونه، باید آن را با تمامی نمونه‌های موجود در حافظه مقایسه و معیار شباهت را محاسبه کرد. پیچیدگی زمانی این الگوریتم در بدترین حالت از مرتبهٔ حاصل ضرب تعداد نمونه‌های مجموعهٔ آموزش در تعداد ویژگی‌های هر نمونه است. بنابراین اندازهٔ پیکرهٔ آموزشی در پیچیدگی زمانی جست‌وجو تأثیر دارد. به منظور کاهش حافظهٔ مورد نیاز برای نگه‌داری نمونه‌ها، می‌توان از یک درخت تصمیم[۲] استفاده کرد [۴].

هر یک از گره‌ها در این درخت، شامل آزمون مقدار برای یک ویژگی خاص و محتمل‌ترین رده برای آن گره است. وزن یال‌ها، مشخص‌کنندهٔ مقدار ویژگی مورد آزمون در گرهٔ مبدأ است. هر مسیر از ریشه به برگ‌های این درخت نیز نشان‌دهندهٔ بردار ویژگی‌های یک نمونهٔ خاص در مجموعهٔ آموزش است. ترتیب ویژگی‌هایی که در گره‌های این درخت مورد آزمون قرار می‌گیرند با استفاده از بهرهٔ اطلاعاتی هر ویژگی به دست می‌آید. بنابراین ابتدا ویژگی‌ای مورد آزمایش قرار می‌گیرد که در رده‌بندی از اهمیت بیش‌تری برخوردار هستند. در نظر گرفتن این ترتیب، سبب می‌شود تمامی گره‌های هم‌سطح، یک ویژگی

---
[۱] Information Gain (IG)
[۲] Decision Tree

واحد را مورد آزمایش قرار دهند و ویژگی مورد آزمایش در گرهٔ ریشه، دارای بیش‌ترین بهرهٔ اطلاعاتی باشد. به همین علت، به این درخت در [۴]، "درخت بهرهٔ اطلاعات[۱]" گفته می‌شود.

برای افزایش سرعت رده‌بندی و کاهش حجم درخت، می‌توان تنها ویژگی‌هایی را در مسیر یک نمونه، مورد آزمایش قرار داد که به رفع ابهام ردهٔ نمونه منجر می‌شوند. به این ترتیب دیگر نیاز نیست تمام ویژگی‌های موجود در بردار ویژگی‌ها، در درخت ظاهر شوند [۴]. به عنوان مثال اگر ردهٔ یک نمونه در سطح سوم درخت تصمیم مشخص شده است، نیازی به ذخیرهٔ سایر ویژگی‌ها در سایر سطوح برای این نمونه و تشکیل یک مسیر کامل نیست. به این ترتیب با هرس کردن بعضی زیردرخت‌ها کارآیی جست‌وجو افزایش یافته، حافظهٔ مورد نیاز برای ذخیرهٔ درخت کاهش می‌یابد.

برای مشخص کردن ردهٔ هر نمونهٔ جدید، جست‌وجو از گرهٔ ریشه آغاز می‌شود. در هر گره، با بررسی مقدار یک ویژگی، رده‌بندی به یکی از گره‌های فرزند هدایت می‌شود. این روند تا رسیدن به گرهٔ برگ ادامه می‌یابد. هر برگ مشخص کنندهٔ یک ردهٔ واحد برای نمونهٔ جدید است. در صورتی‌که بردار ویژگی نمونهٔ جدید، با مسیر طی شده برای هیچ برگی منطبق نشود، الگوریتم عقب‌گرد نموده، به گرهٔ غیرپایانی قبلی بازمی‌گردد. گره‌های غیرپایانی محتمل‌ترین رده را برای مسیری که برای رسیدن به آن‌ها طی شده است، نشان می‌دهند. در چنین مواردی ردهٔ نمونهٔ جدید مبهم خواهد بود.

۴-۴- ساختار سامانهٔ برچسب‌گذاری مبتنی بر حافظه

ساختار یک سامانهٔ برچسب‌گذاری مبتنی بر حافظه از سه قسمت تشکیل می‌شود [۴]:

۱- واژه‌نامه که شامل تمامی واژه‌های موجود در پیکرهٔ آموزشی، برچسب‌های مختلف هر واژه، به همراه فراوانی رخداد هر برچسب است؛

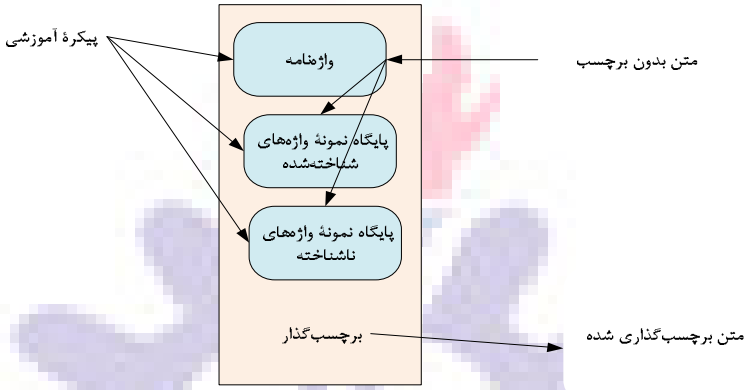
۲- پایگاه نمونه[۲] برای واژه‌های شناخته‌شده؛ و

۳- پایگاه نمونه برای واژه‌های ناشناخته.

---

[۱] IGTree
[۲] Case Base

این سه بخش از پیکرهٔ آموزشی استخراج می‌شوند. برای برچسب‌گذاری هر واژه، ابتدا واژه‌نامه جست‌وجو می‌شود. اگر واژهٔ مورد نظر در واژه نامه موجود باشد، از پایگاه نمونهٔ واژه‌های شناخته شده و در غیر این‌صورت از پایگاه نمونهٔ واژه‌های ناشناخته، برای برچسب‌گذاری استفاده می‌شود. این مراحل در شکل ۴-۱ نشان داده شده‌اند.



شکل ۴-۱- ساختار برچسب‌گذار مبتنی بر حافظه [۴]

۴-۴-۱- ساختار پایگاه نمونهٔ واژه‌های شناخته شده

نمونه‌های موجود در این پایگاه، شامل اطلاعاتی راجع به واژه، زمینهٔ چپ و راست آن واژه و ردهٔ منتسب به واژه در آن بافت هستند. به این رویکرد برای استخراج اطلاعات از پیکرهٔ آموزشی، روش پنجره[۱] گفته می شود [۲۲]. تعداد واژه‌های اطراف، که در برچسب‌گذاری یک واژه مؤثرند، اندازهٔ زمینهٔ واژه نامیده می‌شود. اندازهٔ زمینهٔ چپ و راست در این روش در حقیقت مشخص کنندهٔ اندازهٔ پنجره هستند. اطلاعات موجود در هر نمونه شامل موارد زیر است:

۱- واژه؛

۲- برچسب منتسب به واژه در متن؛

۳- برچسب مبهم زمینهٔ راست؛ و

۴- برچسب ابهام‌زدایی شدهٔ دو واژهٔ سمت چپ.

[۱] Windowing approach

برچسب ابهام‌زدایی شده برای زمینهٔ چپ از نتایج قبلی سامانهٔ برچسب‌گذاری حاصل می‌شود. نتایج آزمایشات در [۴] نشان می‌دهد که اندازهٔ این بافت، به نتایج قابل قبولی در برچسب‌گذاری یک واژهٔ شناخته شده می‌انجامد. برای نگه‌داری نمونه‌ها در حافظه از درخت معرفی شده در بخش ۴-۳ استفاده می‌شود [۴]. س

۴-۴-۲- ساختار پایگاه نمونهٔ واژه‌های ناشناخته

همان‌طور که در فصل سوم اشاره شد در بسیاری از روش‌ها از جمله روش‌های مبتنی بر گذار، برای برچسب‌گذاری واژه‌های ناشناخته از تحلیل ساخت‌واژی آن‌ها استفاده می‌شود. تحلیل ساخت‌واژی این امکان را فراهم می‌کند که واژه‌های ناشناخته را با استفاده از ترکیبات مشخص، به تک‌واژهای[۱] شناخته شده مرتبط کرده و با کمک این ارتباط چند ردهٔ مبهم را برای واژه‌های ناشناخته در نظر بگیریم. سپس با استفاده از اطلاعات زمینهٔ واژه، به رفع ابهام از ردهٔ آن بپردازیم.

در روش مبتنی بر حافظه برای نمایش نمونهٔ یک واژهٔ ناشناخته، هم از ویژگی‌های ساخت‌واژی وهم از زمینهٔ واژه استفاده می‌شود. ویژگی‌های ساخت واژی نمونه شامل موارد زیر می‌شوند [۴]:

- سه حرف پایانی واژه که به صورت سه ویژگی مجزا در بردار ویژگی‌ها نمایش داده می‌شوند. این ویژگی‌ها می‌توانند اطلاعاتی راجع به پس‌وند احتمالی واژه را در اختیار برچسب‌گذار قرار دهند؛ و
- اولین حرف نیز برای تشخیص پیش‌وند، در بردار ویژگی‌های نمونه نشان داده می‌شود. بزرگ و یا کوچک بودن این حرف در زبان انگلیسی می‌تواند به تشخیص عام و یا خاص بودن واژه نیز کمک کند.

اطلاعات بافت، با این فرض که واژه‌های قبلی و بعدی، واژه‌های شناخته شده هستند، شامل دو مورد زیر است:

- برچسب ابهام‌زدایی شدهٔ واژهٔ سمت چپ که نتیجهٔ برچسب گذاری‌های قبلی است؛ و
- برچسب مبهم واژهٔ سمت راست.

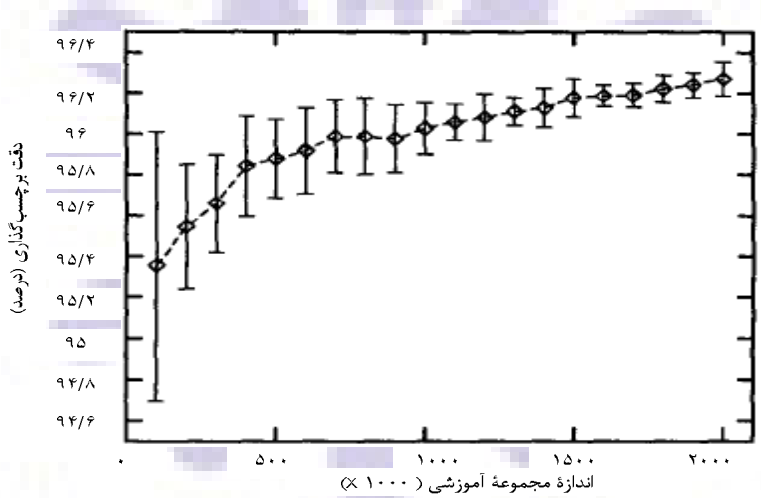
در این پایگاه نیز نمونه‌ها در قالب درخت نگه‌داری می‌شوند. به دلیل بهرهٔ اطلاعاتی بالاتر ویژگی‌های ساخت‌واژی، این ویژگی‌ها در سطوح بالاتر درخت آزمایش می‌شوند. اطلاعات بافت معمولاً در سطوح پایین‌تر و به منظور رفع ابهام مورد استفاده قرار می‌گیرند.

---

[۱] Morphemes

۴-۵- منحنی یادگیری روش مبتنی بر حافظه

شکل ۴-۲ درصد دقت برچسب‌گذاری مبتنی بر حافظه با استفاده از درخت بهرهٔ اطلاعات را بر حسب اندازهٔ پیکرهٔ آموزشی مورد استفاده نشان می‌دهد. پیکرهٔ مورد استفاده در این آزمایش پیکرهٔ روزنامهٔ وال استریت[۱] [۲۳] است. هر یک از آزمایشات با استفاده از روش اعتبارسنجی متقابل ده قسمتی[۲]، انجام شده است. در این روش مجموعهٔ داده‌ها ده بار به دو بخش آموزش و آزمون تقسیم می‌شود. در هر یک از این ده آزمون، ۹۰ درصد مجموعهٔ داده‌ها به عنوان مجموعهٔ آموزش و ۱۰ درصد باقیمانده به عنوان مجموعهٔ آزمون در نظرگرفته می‌شود. دقت کلی آزمایش از میانگین‌گیری دقت گزارش شده در آزمون‌ها به دست می‌آید.



شکل ۴-۲- منحنی یادگیری روش مبتنی بر گذار [۴]

در منحنی یادگیری شکل ۴-۲ مشاهده می‌شود که با افزایش اندازهٔ پیکرهٔ آموزشی، دقت برچسب‌گذاری این روش نیز افزایش می‌یابد. در شکل ۴-۲ محدودهٔ تغییر دقت برچسب‌گذاری در هر یک از آزمایشات نشان داده شده است. با افزایش اندازهٔ مجموعهٔ داده، کران بالا و پایین این تغییرات به یکدیگر نزدیک‌تر می‌شوند. این مسأله نشان می‌دهد که هر چه اندازهٔ مجموعهٔ آموزش (و یا مجموعهٔ داده) بزرگ‌تر شود، میزان تغییر دقت روش در تشخیص برچسب‌ها کم‌تر می‌شود [۴].

---

[۱] Wall Street Journal (WSJ)
[۲] 10 Fold Cross Validation

۴-۶- نتیجه‌گیری

روش برچسب‌گذاری مبتنی بر حافظه در مقایسه با سایر روش‌های مطرح شده در این پژوهش‌نامه مثل روش مدل مخفی مارکوف و روش مبتنی بر گذار مزایایی دارد، از جمله:

۱- این روش با استفاده از پیکره‌های آموزشی کوچک نیز به دقت نسبتاً بالایی در برچسب‌گذاری می‌رسد و به پیکره‌های خیلی بزرگ برای یادگیری نیاز ندارد [۴].

۲- دقت روش مبتنی بر حافظه در برچسب‌گذاری واژه‌های ناشناخته، بیشتر از سایر روش‌هاست. علاوه بر این نکته، در بسیاری از روش‌ها که برای برچسب‌گذاری واژه‌های ناشناخته از تحلیل ساخت‌واژی استفاده می‌کنند، منابع دیگری مثل قوانین املایی و ساخت‌واژی زبان و مجموعه‌ای شامل تکواژهای زبان نیاز است. این امر سبب می‌شود که سامانهٔ برچسب‌گذاری وابستگی اساسی به زبان داشته باشد. در حالی‌که در برچسب‌گذاری مبتنی بر حافظه، چنین وابستگی به زبان مبدأ وجود ندارد [۴].

۳- نتایج آزمایشات در [۴] نشان می‌دهد پیچیدگی زمانی جست‌وجو برای ردهٔ یک نمونهٔ جدید در درخت، متناسب با رابطهٔ (۴-۴) است. در این رابطه F تعداد ویژگی‌ها و V میانگین تعداد مقادیری است که برای هر ویژگی وجود دارد. از آن‌جایی‌که پیچیدگی زمانی رده‌بندی با استفاده از درخت، به اندازهٔ مجموعهٔ آموزش وابسته نیست، این روش برای پیکره‌های آموزشی بزرگ مناسب است. در صورت نگه‌داری نمونه‌ها بدون استفاده از درخت، مرتبهٔ زمانی رده‌بندی، به اندازهٔ پیکرهٔ آموزشی نیز مرتبط بود.

$$\theta = F * \log V \qquad (۴-۴)$$

فصل پنجم

بررسی روش‌های برچسب‌گذاری در زبان فارسی

۵-۱- بررسی بخش‌های سامانهٔ برچسب‌گذاری متون فارسی

از هر یک از روش‌هایی که در فصل‌های گذشته به آن‌ها اشاره شد، می‌توان برای برچسب‌گذاری متون فارسی نیز استفاده کرد. ولی تفاوت‌های زبان فارسی با زبان‌های دیگر سبب ایجاد تفاوت‌هایی در طراحی یک سامانهٔ برچسب‌گذاری برای زبان فارسی با سایر سامانه‌های برچسب‌گذاری می‌شود. در سامانه‌های فارسی، برای افزایش دقت، بسیاری از مراحل پردازشی قبل و یا بعد از بخش مرکزی برچسب‌گذار انجام می‌شوند؛ منظور از بخش مرکزی سامانهٔ برچسب‌گذاری، بخشی است که در آن با استفاده از یکی از روش‌ها، به برچسب‌گذاری واژه‌ها و جملات ورودی پرداخته می‌شود. به عنوان نمونه، در سامانه‌ای که در [۵] معرفی شده است، پیش از برچسب‌گذاری با روش‌های دونگاشت، سه‌نگاشت و روش مبتنی بر حافظه، بخش‌های زیر وجود دارند:

- بخش تشخیص‌دهندهٔ کران جمله و واژه: این مسأله در سایر زبان‌ها نیز وجود دارد ولی تشخیص کران واژه در زبان فارسی یکی از چالش‌های سامانه‌های برچسب‌گذاری است. یکی از علل اصلی این مشکل وجود فاصله بین اجزای یک واژه است که حتی در بسیاری از نوشته‌های رسمی، نیز شکل درست آن رعایت نمی‌شود. به عنوان نمونه، این نکته که در یک سامانه، "کم تر" یک واژه در نظر گرفته شود (صفت تفضیلی) و یا دو واژه که متشکل از "کم" و "تر" در معنای خیس باشد. از آن‌جایی که این مسأله بر فراوانی یک برچسب خاص در پیکره و یا بافت واژه‌های اطرافش تأثیرگذار است، از اهمیت بالایی برخوردار است.

• تحلیل‌گر ساخت‌واژی: در صورتی‌که واژه‌ای در پیکرهٔ آموزشی یافت نشود به جای اینکه مستقیماً از روش‌های برچسب‌گذاری واژه‌های ناشناخته در مورد آن استفاده شود، ابتدا به کمک یک تحلیل‌گر ساخت‌واژی مورد بررسی قرار می‌گیرد. در زبان فارسی به علت وجود ساخت‌واژه‌های تصریفی و اشتقاقی احتمال ساخت واژه‌های جدیدی که در پیکرهٔ آموزشی وجود ندارند زیاد است. این تحلیل‌گر کمک می‌کند تا بتوان چنین واژه‌هایی را شناسایی نموده و سپس برای برچسب‌گذاری آن‌ها اقدام کرد. البته تحلیل‌گرهای موجود، تنها قادر به تشخیص ساخت‌واژه‌های تصریفی هستند و با آن‌ها نمی‌توان ساخت‌های اشتقاقی را تشخیص داد.

• ابهام‌زدایی از هم‌نگاره‌ها: در این مرحله هم‌نگاره‌بودن واژهٔ ورودی بررسی می‌شود. در صورتی که این واژه یکی از هم‌نگاره‌های موجود در پیکرهٔ آموزشی باشد، ابهام‌زدایی می‌شود.

از جمله مشکلاتی که در سامانه‌های برچسب گذاری فارسی وجود دارد، وجود بعضی جملات و یا عبارت‌های عربی در پیکرهٔ آموزشی است [۵]. این مشکل در پیکرهٔ بی‌جن‌خان نیز وجود دارد زیرا این پیکره مجموعه‌ای از متون روزنامهٔ هم‌شهری است و وجود چنین عباراتی در آن محتمل است. در صورت برخورد با چنین عباراتی، برچسب‌گذارهای آماری نمی‌توانند ارتباط بین اجزاء جمله را به درستی تشخیص دهند. از طرفی به علت مشترک‌بودن بسیاری از واژه‌های عربی و فارسی مثل "قبل" یا "من" ممکن است برچسب نادرست به آن‌ها منتسب شود. اگر برخی از این واژه‌های عربی در مجموعهٔ واژه‌های ناشناخته قرار گیرند، روش‌هایی که تا کنون در تشخیص برچسب واژه‌های ناشناخته معرفی شدند قادر به برچسب‌گذاری این واژه‌ها نخواهند بود. علت آن است که ساختار واژه‌های عربی با ساختار واژه‌های فارسی متفاوت است. همین عامل دقت برچسب‌گذاری را کاهش می‌دهد. از این رو یکی از اقداماتی که در راستای افزایش دقت سامانه‌های فارسی می‌توان انجام داد در نظر گرفتن بخشی برای تشخیص عبارات عربی است.

یکی از نکاتی که باید در سامانه‌های برچسب‌گذاری فارسی مورد توجه قرار داد، عبارات و جملات غیر رسمی و عامیانه هستند. بسیاری از عبارات و واژه‌های عامیانه، که در زبان محاوره نیز کاربرد دارند، شکسته شده و از ساخت‌واژه‌های فارسی برای تحلیل آن‌ها نمی‌توان استفاده کرد. از طرفی در بسیاری از جملات غیر رسمی، ساختار معمول جمله‌های فارسی دیده نمی‌شود، از این رو اطلاعات بافت واژه‌ها نیز در برچسب‌گذاری کمک زیادی نمی‌کند. حتی ممکن است به انتساب برچسب

نادرست نیز بینجامد. بسیاری از روش‌ها مثل روش مبتنی بر حافظه برای انتساب برچسب، به بافت واژه توجه ویژه دارند. این روش‌ها در صورت برخورد با این عبارات، دقت بالایی در برچسب‌گذاری نخواهند داشت.

بسیاری از روش‌هایی که در فصل‌های گذشته معرفی شدند مانند روش‌های مبتنی بر گذار، روش‌های مبتنی بر حافظه و یا روش‌هایی که از مدل پنهان مارکوف استفاده می‌کنند، برای برچسب‌گذاری متون به یک پیکرهٔ برچسب‌خورده نیاز دارند. از این رو وجود پیکره‌ای با حجم مناسب در دقت این روش‌ها تأثیرگذار است. برخی نکاتی که در ایجاد و برچسب‌گذاری یک پیکرهٔ متنی باید مورد توجه واقع شود، عبارتند از:

- فراوانی انواع مختلف هم‌نگاره‌ها در پیکره؛
- مجموعهٔ برچسب‌های در نظر گرفته شده برای برچسب‌گذاری پیکرهٔ آموزشی؛ و
- یکپارچگی متن در استفاده از اصول نگارش فارسی مانند استفاده از فاصلهٔ کامل و یا نیم‌فاصله.

۵-۲- بررسی نتایج سه روش برچسب‌گذاری در متون فارسی

در [۱۶] نتایج برچسب‌گذاری پیکرهٔ بی‌جن‌خان با استفاده از سه روش مبتنی بر حافظه، تخمین احتمال بیشینه و مدل پنهان مارکوف بررسی شده است. در سومین روش از مدل مارکوف درجهٔ دو استفاده می‌شود. در یک آزمایش، دقت برچسب‌گذاری تمام این روش ها بدون اعمال پردازش‌های جانبی که در ۵-۱ به آن‌ها اشاره شد، مورد بررسی قرار گرفته است. همان‌طور که در جدول ۵-۱ مشاهده می‌شود، در این آزمایش دقت برچسب‌گذاری روش مبتنی بر مدل پنهان مارکوف در بخش واژه‌های ناشناخته بیش‌تر از سایر روش‌ها است. دقت نهایی این روش نیز از دو روش دیگر بیش‌تر است.

جدول ۵-۱- دقت برچسب‌گذاری بدون استفاده از قوانین ساخت‌واژی در برچسب‌گذاری واژه‌های ناشناخته (برحسب درصد) [۱۶]

روش برچسب‌گذاری	نهایی	واژه های ناشناخته	واژه‌های شناخته شده
مدل مارکوف درجهٔ دو	۹۶/ ۶۴	۷۷/۷۷	۹۷/۰۱
مبتنی بر حافظه	۹۶/۴۲	۷۵/۱۵	۹۶/۸۶
تخمین احتمال بیشینه	۹۴/۶۳	۱۵	۹۶/۶۰

در این مرحله برای تشخیص برچسب واژه‌های ناشناخته به ساخت‌واژهٔ آن ها هیچ توجهی نمی‌شود و صرفاً از روش‌های آماری برای برچسب‌گذاری استفاده می‌شود. نمونه‌ای از این روش‌ها نسبت دادن برچسب با بیش‌ترین فراوانی به واژه‌های ناشناخته است. در آزمایش دیگری برای برچسب‌گذاری واژه‌های ناشناخته در روش‌های مبتنی بر حافظه و تخمین احتمال بیشینه، از تحلیل ساخت‌واژی (تشخیص برخی وندها در واژه) نیز استفاده می‌شود. اگرچه تحلیل‌گر مورد استفاده بسیار ساده است و با تعداد قوانین اندکی ساخت‌واژه‌ها را بررسی می‌کند ولی جدول ۵-۲ افزایش دقت این روش‌ها را نشان می‌دهد.

جدول ۵-۱- دقت برچسب‌گذاری با استفاده از قوانین ساخت‌واژی در برچسب‌گذاری واژه‌های ناشناخته (برحسب درصد) [۱۶]

روش برچسب‌گذاری	نهایی	واژه های ناشناخته	واژه‌های شناخته شده
مدل مارکوف درجهٔ دو	۹۶/۶۴	۷۷/۷۷	۹۷/۰۱
مبتنی بر حافظه	۹۶/۶۳	۸۱/۱۱	۹۶/۸۶
تخمین احتمال بیشینه	۹۵/۹۷	۶۵/۷۵	۹۶/۶۰

بنابراین می‌توان نتیجه‌گیری کرد در صورت افزایش حجم تحلیل‌گر ساخت‌واژی، میزان دقت این روش‌ها نیز افزایش یابد.

۵-۳- کلمات ناشناخته در فارسی

در فصل سوم پژوهش‌نامه به یکی از روش‌های برچسب‌گذاری واژه‌های ناشناخته که در برخی برچسب‌گذارهای مبتنی بر گذار نیز استفاده می‌شود، اشاره شد. در این روش به واژه‌های ناشناخته با توجه به کوچک و یا بزرگ بودن حرف اولشان، برچسب "اسم عام" یا "اسم خاص" منتسب می‌شود. سپس با استفاده از یک الگوریتم یادگیری، برچسب واژه تصحیح می‌شود [۳]. از آن‌جایی که در زبان فارسی، اسامی خاص و عام تفاوت نوشتاری ندارند. روش دیگری برای برچسب‌گذاری واژه‌های ناشناخته استفاده می‌شود. در [۱۶]، برای برچسب‌گذاری واژه‌های ناشناخته از پرکاربردترین برچسب در مجموعهٔ آموزشی استفاده می‌شود. در بخشی از پیکرهٔ بی‌جن‌خان که در این آزمایش به عنوان مجموعهٔ آموزشی استفاده شده است، برچسب "اسم مفرد" پرکاربردترین برچسب بوده است. اگرچه دقت برچسب‌گذاری واژه‌های ناشناخته با این روش در زبان فارسی، از دقت به دست آمده در سایر زبان‌ها مانند انگلیسی و آلمانی کم‌تر است، ولی نتایج تفاوت زیادی را در دقت برچسب‌گذاری نشان نمی‌دهند [۱۶].

۵-۴- کارهای آینده

با توجه به نکاتی که در بخش ۵-۱ به آن‌ها اشاره شد، می‌توان رئوس فعالیت‌های آینده در زمینهٔ برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی را به این صورت برشمرد:

- تشخیص عبارات و جملات عربی در بین متون فارسی؛
- تشخیص واژه‌ها و عبارات غیررسمی و محاوره‌ای در متون فارسی و بررسی رفتار آن‌ها؛
- بررسی واژه‌های ناشناخته در زبان فارسی و یافتن روش‌هایی برای برچسب‌گذاری دقیق آن‌ها.

منابع

1. E. Charniak, C. Hendrickson, N. Jacobson and M. Perkowitz, “Equation for part-of-speech tagging”, Proceedings of the Eleventh National Conference on Artificial Intelligence, 1993, pp. 784-789.
2. A. Ratenaparkhi, “A Maximum Entropy Model for Part-Of-Speech Tagging”, Proceeding of the Conference on Empirical Methods in Natural Language Processing, 1996, pp. 133-142.
3. E. Brill , ”Transformation-Based Error-Driven Learning and Natural Language
    a. Processing: A Case Study in Part-of-Speech Tagging”, Computational Linguistic, 21(4), 1995, pp. 543-565.
4. W. Daelemans, J. Zavral, P. Berck and S. Gillis, “MBT: A Memory-Based Part of Speech Tagger-Generator”, Proceeding of the Fourth Workshop on Very Large Corpora, 1996, pp 14-27.

۵. م. محسنی، "سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکرهٔ متنی زبان فارسی"، پایان‌نامهٔ کارشناسی ارشد، دانشگاه علم و صنعت ایران، دانشکدهٔ مهندسی کامپیوتر، ۱۳۸۷.

6. D. Jurafsky and M. James, “Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics”, Prentice Hall, 2000, pp. 286-308.
7. S. Atkins, J. H. Clear and N. Ostler, “Corpus Design Criteria”, Literary & Linguistic Computing, Vol. 7, No. 1, 1992, pp. 1-16.

۸. م. بی‌جن‌خان، "امکان‌سنجی برای طرح مدل‌سازی زبان فارسی"، مجله علمی پژوهشی دانشکده ادبیات و علوم انسانی دانشگاه تهران، شماره ۱۶۳-۱۶۲، دوره ۵۰-۵۱، صص. ۸۱-۹۶.

9. S. Klein and R. F. Simmons, “A Computational Approach to Grammatical Coding of English Words”, Journal of the Association for Computing Machinery, 10(3), 1963, pp. 334–347.
10. C. Manning and H. Schutze, “Foundations of Statistical Natural Language Processing”, MIT Press, 1999.
11. S. Russell and P. Norvig, “Artificial Intelligence: A Modern Approach”, Prentice Hall, 2003, pp. 649-763.
12. K. Megerdoomian, “Developing a Persian Part Of Speech Tagger”, In Proceedings of First Workshop on Persian Language and Computers, Tehran University, Iran, 2004.

۱۳. م. بی‌جن‌خان و ش. مرادزاده، " هم‌نگاره‌های خط فارسی"، مجموعه سخنرانی‌ها و گزارش‌ها و چکیدهٔ طرح‌ها، اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشکده ادبیات و علوم انسانی دانشگاه تهران، ۱۳۸۳، صص ۵۳-۶۳.

14. S. M. Assi and M. Haji Abdolhoseini, “Grammatical Tagging of a Persian Corpus”, International Journal of Corpus Linguistics, Vol. 5, Number 1, 2000, pp. 69-81.
15. H. Schutze, “Distributional Part-of-Speech Tagging”, Online Proceedings of the ACL SIDGAT Workshop, 1995.
16. F. Raja, H. Amiri, S. Tasharofi, M. Sarmadi, H. Hojjat and F. Oroumchian, “Evaluation of Part of Speech Tagging on Persian Text”, Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages Linguistic Institute, Stanford, California, USA, 2007.
17. D. Yarowsky, “Homograph Disambiguation in Text-to-speech Synthesis”, Progress in Speech syntheses, Springer-Verlag, 1996, pp. 159-175.
18. L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proceedings of the IEEE 77(2), 1989, pp. 257–286.
19. C. Samuelsson and A. Voutilainen, “Comparing a Linguistic and a Stochastic Tagger”, Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Madrid, 1997, pp. 246-253.
20. A. Voutilainen and J. Heikkila, “An English constraint grammar (**ENGCG**): a surface-syntactic parser of English”, Research Unit for Computational Linguistics University of Helsinki, 1993, pp. 189-199.
21. E. Dematas and G. Kokkinakis, “Automatic Stochastic Tagging of Natural Language Texts”, Computational Linguistics, 21(2), 1995.
22. T. Sejnowski and C. Rosenberg, “Parallel networks that learn to pronounce English text”, Complex Systems 1, 1987, pp. 145-168.
23. D. P. Paul, J. M. Baker, “The Design for the Wall Street Journal-based CSR Corpus”, Human Language Technology Conference (HLT), Proceedings of the workshop on Speech and Natural Language, 1992, pp. 357-362.