تحلیل ساختار جملات انگلیسی با استفاده از یادگیری ماشین

چکیده

زبان‌های طبیعی به هرکدام از زبان‌هایی گفته می‌شود که توسط بشر در دنیا صحبت می‌شود و خودبه‌خود پدید آمده و تکامل‌یافته است. پردازش زبان‌های طبیعی یکی از زیرشاخه‌های بااهمیت در حوزه ی گسترده علوم کامپیوتر، هوش مصنوعی و نیز دانش زبان شناسی محاسباتی است که به تعامل بین کامپیوتر و زبان‌های انسانی می‌پردازد. چالش اصلی و عمده در زمینه پردازش زبان طبیعی درک زبان طبیعی و ماشینی کردن فرایند درک و برداشت مفاهیم بیان شده با یک زبان طبیعی انسانی است. هدف اصلی در پردازش زبان طبیعی، ایجاد تئوری‌هایی محاسباتی از زبان، با استفاده از الگوریتم‌ها و ساختارهای داده‌ای موجود در علوم کامپیوتر است. بدیهی است که در راستای تحقق این هدف، نیاز به دانشی وسیع از زبان است و علاوه بر محققان علوم کامپیوتر، نیاز به دانش زبان شناسان نیز در این حوزه می باشد. با پردازش اطلاعات زبانی می‌توان آمار موردنیاز برای کار با زبان طبیعی را استخراج کرد.

هدف اولیه هر برنامه NLP ایجاد یک درخت تجزیه برای یک جمله متعلق به مجموعه آن زبان است. برای طبقه‌بندی صحیح کلمات که به کدام نوع خاص تعلق دارد به مدل زبان تکیه می‌شود. برای مشخص کردن این طبقه‌بندی از الگوریتم‌های یادگیری ماشین استفاده می‌شود ماشین یادگیر با استفاده از الگوریتم‌های مخصوص شروع به یادگیری جایگاه و نوع کلمات می‌نماید و سپس برای جملات جدید با استفاده از یادگیری خود اقدام به طبقه‌بندی می‌نماید. با مشخص شدن نوع جملات با استفاده از مدل‌های منطقی می‌توان درخت تجزیه جملات را رسم کرد در این پرویه از مجموعه‌ای شامل نمونه‌های بیش از ۱۰۰۰۰۰ جمله از جملات زبان انگلیسی است که با استفاده از الگوریتم‌های یادگیری آموزش داده‌شده است استفاده شده است و با استفاده از زبان برنامه‌نویسی C# درخت تجزیه جملات به‌درستی به‌دست‌آمده است.

۱-۱ مقدمه

هدف اصلی NLP درک خودکار از زبان نیمه ساختاریافته‌ای است که انسان‌ها از آن استفاده می‌کنند. این مطالعه در زمینه‌های گوناگون مانند تجزیه‌وتحلیل معنایی، خلاصه، طبقه‌بندی متن و غیره کاربرد دارد. در مقایسه با حوزه‌هایی مانند مطالعه الگوریتم‌ها که به‌خوبی شناخته‌شده‌اند، NLP هنوز در حال ظهور است و تحقیقات زیادی در این مسیر وجود دارد. NLP به‌شدت وابسته موضوعات پایه مانند آمار، نظریه احتمالات و تئوری محاسبات است ولی قبل از پرداختن به این خصوصیات ابتدا بایستی در حوزه زبان‌شناختی و آشنایی با نحو، معانی و درک زبان‌ها تحقیقات انجام شود. در این فصل ابتدا به موضوع زبان‌شناسی پرداخته می‌شود سپس مروری بر تحقیقات انجام‌شده در زمینه پردازش زبانه‌ای طبیعی ارائه می‌شود.

۱-۲ زبان‌شناسی

زبان قاعده‌مندترین علم در حوزه‌ی علوم انسانی است آن‌چنان‌که نظم دقیق آن را تنها با علم ریاضی می‌توان سنجید. قاعده‌مندی و نظم زبان آن‌چنان است که هیچ تخلفی از اصول آن جایز شمرده نمی‌شود. می‌دانیم که بنیاد زبان بر محور هم‌نشینی استوار است و بیشترین بهره را از زبان برای ایجاد ارتباط و انتقال تجربیات می‌بریم. پیشرفت علوم گوناگون در جوامع بشری حاصل همین استفاده از زبان است و بهترین تعریف‌هایی که تاکنون از زبان شده در ارتباط با همین کارکرد زبان است. زبانشناسی امروز علمی است مانند همه‌ی علوم دیگر، دارای اصول و قواعد و روش‌های مبتنی بر مشاهده و تجربه و استقرا، بر اساس ملاک‌های صوری و عینی. قطعیت علمی بسیاری از دستاوردهای زبانشناسی در قرن حاضر آن را به مرز علوم دقیق نزدیک کرده است. در زبانشناسی امروز از میان دیدگاه‌هایی که در مورد زبانشناسی وجود دارد می‌توان از زبانشناسی ساختگرا و زبانشناسی گشتاری-زایایی و نام برد [٤].

۱-۲-۱ زبانشناسی ساختگرا

در این حوزه درباره‌ی ساخت جمله می‌توان آن را به عناصر کوچک‌تری تجزیه نمود تا سرانجام واحدهایی به دست آید که دیگر قابل‌تجزیه نباشند، یعنی اگر آن‌ها بیشتر تجزیه کنیم اجزایی به دست می‌آید که دیگر نه معنی دارد و نه می‌توانند یک نقش دستوری بگیرند؛ که ما به آن تکواژ می-گوییم. البته اکثر تکواژها را به عناصر ریزتری می‌توان تجزیه کرد که آن‌ها واحدهای صوتی هستند که نه معنی می‌دهد و نه نقش دستوری. برای نامیدن این واحدهای صوتی اصطلاح phoneme به‌کاررفته که معادل آن‌هم در فارسی واج قرارگرفته است [۷].

۱-۲-۲ زبان‌شناسی گشتاری -زایایی

نوام چامسکی، زبانشناسی که امروز شهرتی جهانی دارد. کتاب ساختارهای نحوی خود را در سال ۱۹۵۷منتشر ساخت. این کتاب برای اولین بار مهم‌ترین و انسجام بافته‌ترین نظریه‌ی زبانی را در تاریخ زبانشناسی به وجود آمده بود، یعنی دستور زایا-گشتاری را به جهانیان معرفی نمود. دستور زایا- گشتاری نظریه‌ای بود سخت انقلابی که دچار تحولات زیادی شده است، اما جهت این تحول دور شدن هر چه بیشتر از نظریات ساختگرایان بوده است. این نظریه، به‌طوری‌که از نام آن برمی‌آید، از دو جنبه‌ی متمایز ولی مربوط به هم تشکیل‌شده است. چنانچه قبلاً گفته شد، دستور ساختاری می‌کوشد جمله‌های زبان را به عناصر سازنده یا سازه‌های ریزتر تجزیه کند و آن‌ها را در رابطه باهم طبقه‌بندی نماید. چامسکی این خصوصیت را وجه اشتراک همه‌ی نظریه‌های ساختاری می‌شمارد و همه‌ی آن‌ها را علیرغم تفاوت‌هایی که باهم دارند، در یک طبقه قرار می‌دهد. چامسکی این‌طور استدلال می‌کند که اگر قرار باشد دستور زبان بتواند به‌خوبی از عهده‌ی توصیف واقعیات زبانی برآید و بتواند روابط بین جمله‌های زبان را توجیه کند این کافی نخواهد بود که فقط به نشانه‌ها و روابط آشکار و عینی بپردازد بلکه باید به کشف روابط و تفاوت‌هایی که در زیربنای جمله‌های عینی وجود دارد توجه کند. ازاین‌رو او برای هر جمله‌ای دو نوع ساخت قائل می‌شود: یکی

ژرف‌ساخت که درواقع تعیین‌کننده‌ی روابط معنایی و منطقی اجزای جمله است و دیگری روساخت که شکل خارجی و عینی جمله را نشان می‌دهد و الزاماً منطبق با ژرف‌ساخت جمله نیست. از سوی دیگر معتقد است که ژرف‌ساخت جمله از راه تعداد محدودی قاعده که آن‌ها را قواعد گشتاری می‌نامد به روساخت تبدیل می‌شود. قواعد گشتاری از راه حذف، تعویض، افزایش یا جابه‌جایی روابط ژرف‌ساختی را به روابط روساختی تبدیل می‌نماید. به هریک از این فعل‌وانفعالات یا به مجموعه‌ای از آن‌ها که به کمک یک قاعده صورت می‌گیرد، گشتار گفته می‌شود [۱۹].

۱-۲-۳ کاربردهای عملی پردازش زبان

کاربردهای عملی پردازش زبان توسط کامپیوتر را می‌توان تا حدودی به موارد زیر خلاصه نمود [۲۰]:

- استفاده از زبان آدمی برای فرمان دادن به کامپیوتر به‌جای زبان‌های مصنوعی که هم‌اکنون مورداستفاده قرار می‌گیرند.
- بانک اطلاعاتی و سیستم ارائه‌دهنده کمک به کاربران در کامپیوتر که سوا لاتی را به زبان آدمی از کاربر بگیرد و پاسخ ارائه دهد.
- ترجمه خودکار متون علمی، فنی و تجاری از زبانی به زبان دیگر.
- ساخت خودکار بانک اطلاعاتی از متونی که دارای ماهیت فنی هستند همچون گزارش مربوط به تعمیر دستگاه‌های مختلف یا گزارش سوابق پزشکی بیمار.

تمامی این کاربردها فعلاً به‌صورت مختصر به زبان انگلیسی و سایر زبان‌های اروپایی موجود است. برخی از آن‌ها هم‌اکنون استفاده تجاری نیز پیداکرده‌اند برای مثال ترجمه نتایج جستجوهای مختلف در اینترنت.این برنامه‌ها تاکنون دارای موفقیت نسبی هم بوده‌اند چراکه کامپیوتر جهت انجام آن‌ها الزاماً نیازی

به داشتن اطلاعاتی در مورد دنیای خارج ندارد. برنامه بانک اطلاعاتی، از زبان آدمی جهت ارائه اطلاعات کمک می‌گیرد. برنامه مترجم از این حقیقت سود می‌جوید که متون تخصصی بندرت از حوزه خاص خود خارج می‌شوند. قادرسازی کامپیوتر جهت درک شعر، داستان‌های تخیلی و طنز بسیار مشکل‌تر است چراکه درک چنین مطالبی، مستلزم داشتن تجربیات بسیار زیاد بشری و دانش دنیای پیرامون است. پس پردازش زبان آدمی توسط کامپیوتر، مستلزم تحمیل قیدوبند بر دانش عمومی و تجربیات آدمی است.

۱-۲-٤ نحو و جمله‌سازی

نحو یا جمله‌سازی خلاق‌ترین سطح در زبان آدمی است. آدمی به‌ندرت صدا یا کلمه جدید ابداع می‌کند ولی هر فردی که به یک‌زبان زنده دنیا تکلم می‌کند، دائم در حال خلق جملاتی است که هرگز نه دیده و نه شنیده است. این خلاقیت بدین معنی است که نحو با آواشناسی و واژه-شناسی متفاوت است. روشی خوب جهت توصیف صداها یا فرآیندهای کلمه‌سازی در زبان، تهیه فهرستی از آن است. ولی هیچ‌گاه نمی‌توان فهرست تمامی جملات زبان را ثبت و ضبط کرد چراکه تعداد آن‌ها بسیار زیاد است. در حقیقت اگر حدی را برای طول جمله قائل نشویم، تعداد جملات زبان بینهایت خواهد بود [۱۷]. چامسکی اولین فردی بود که چنین ادعایی کرد. وی دستوری زایشی را مطرح کرد که بجای اینکه مستقیماً جمله یا ساختار آن را فهرست کند، ساخت جمله را با کمک ارائه قانون آن توصیف می‌کند. برای مثال قوانین:

جمله(S) = عبارت اسمی(NP) + عبارت فعلی(VP)

NP = N (اسم)

VP = NP + O +PP (عبارت حرف‌اضافه) +V

PP = P +NP

با استفاده از این قوانین می‌توان جملات زیادی را تولید کند. قوانین ازاین‌دست، علاوه بر زبانشناسی، در علوم کامپیوتری بخصوص طراحی مفسرین جهت پردازش زبان‌های طبیعی به‌صورت استاندارد پذیرفته‌شده‌اند. تشخیص جمله توسط کامپیوتر را پردازش گویند. جهت پردازش جمله، کامپیوتر ساختار جمله را با ساختارهای موجود در حافظه خود می‌سنجد [۱۸]. این روش را هم از بالا به پائین به بالا می‌توان انجام داد. پردازشگر بالا به پائین، ابتدا به دنبال جمله می‌گردد، سپس به قوانین می‌نگرد تا ببینید جمله حاوی چه عناصری است. پردازشگر پائین به بالا ابتدا از کلمات شروع می‌کند و سپس به قوانین می‌نگرد تا ببیند چگونه کلمات به همدیگر مرتبط شده‌اند. بهترین پردازشگر، پردازشگری است که بتواند از هر دو روش استفاده نماید. پردازش زبان انگلیسی به‌طور وسیعی موردمطالعه قرارگرفته است و بسیاری از محققین علوم پردازش زبان طبیعی، آن را تمام‌شده می‌پندارند. نکته فعلی در جوامع انگلیسی‌زبان، این است که بهترین روش کدام است، نه اینکه آیا راهی برای انجام این کار وجود دارد. پردازش زبانه‌ای دیگر خیلی به‌اندازه انگلیسی مطالعه نشده است و در مورد زبان فارسی کارهای خیلی زیادی انجام‌نشده است. بسیاری از فن‌های امروزی پردازش، مبتنی بر ترتیب منظم کلمات هستند چراکه انگلیسی به ترتیب کلمات بسیار حساس است و برای زبان‌هایی همچون فارسی که ترتیب کلمات خیلی مهم نیست خیلی مناسب نیستند. لذا نیاز به کار جدی در این زمینه کاملاً احساس می‌شود [۱۹].

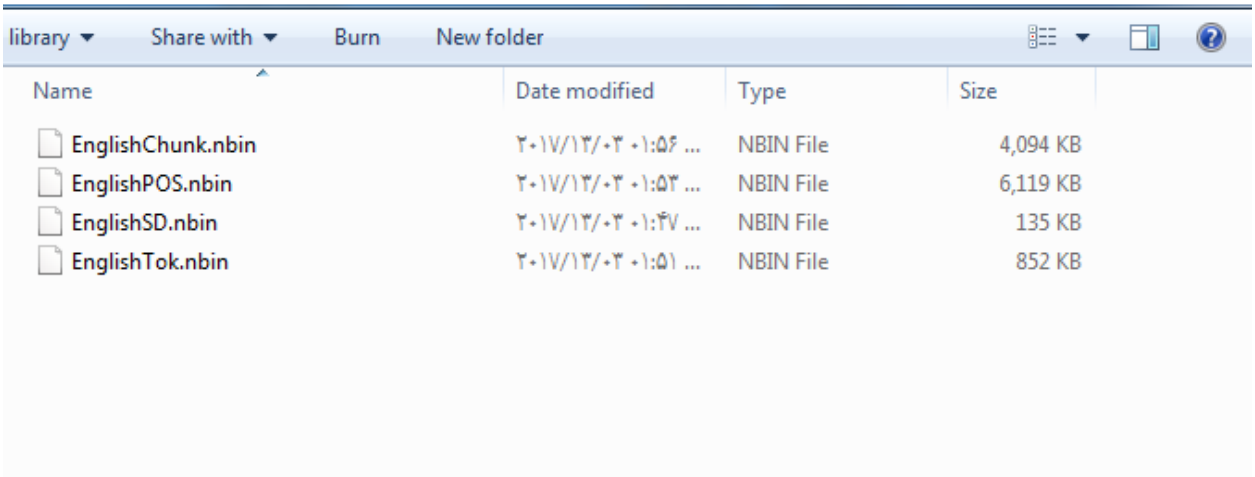
.

۱-۳ مقدمه

در این فصل روش کلی خود را برای پیاده‌سازی تجزیه و پیدا کردن درخت جملات با استفاده از مخازن یاد گیر و نرم‌افزار برنامه‌نویسی C# توضیح خواهیم داد و مراحل هر یک از روش‌ها را به‌طور کامل توضیح می‌دهیم.

۱-٤ پیاده‌سازی ساختار جملات با استفاده از C#

برای پیاده‌سازی با استفاده از C# احتیاج به توابع کتاب‌خانه‌ای تعیین جملات، تعیین کلمات، عبارت‌ها و ... داریم که این توابع با استفاده از الگوریتم‌های یادگیری گردآوری‌شده است و به زبان Java در سایت [۲٤] موجود است این فایل‌ها طبق شکل (۱-٤) عبارت‌اند از:



شکل .Error! No text of specified style in document-۱ فایل‌های مربوط به مخازن موجود به زبان Java

محتویات این فایل‌ها به‌صورت کد جاوا است و شامل نمونه‌های بیش از ۱۰۰۰۰۰ جمله از جملات زبان انگلیسی است که با استفاده از الگوریتم‌های یادگیری آموزش داده‌شده است. جدول ۱-٤ اختصارات مورداستفاده در این گزارش را نشان می‌دهد.

جدول .Error! No text of specified style in document-۱ اختصارات

Abbreviations	Abbreviations Meaning
S	Sentence
Det	Determiner
Adj	Adjective
Pron	Pronoun
Num	Numerals
Conj	Conjunction
Neg	Negation
Prep	Preposition
Adv	Adverb
V	Verb
VC	Verb Command
N	Noun
NP	Noun Phrase
VP	Verb Phrase
AP	Adjective Phrase
NPP	Noun Preposition Phrase
VPP	Verb Preposition Phrase
APP	Adjective Preposition Phrase

برای بررسی کار گرد الگوریتم پیشنهادی در C# ابتدا بایـد اختصـارات اسـتفاده‌شـده در برنامـه نیـز تشریح شوند (به‌عنوان‌مثال برای ضمایر چند نوع ضمیر داریم) که جدول ٤-۲ نشان‌دهنـده ایـن اختصـارات هستند.

Error! No text of specified style in document.-۲ نشانه‌ها و معانی آن‌هاجدول

نشانه	شرح انگلیسی	شرح فارسی
NP	Noun Phrase	عبارت اسمی
VP	Verb Phrase	عبارت فعلی
CC	Coordinating conjunction	حروف ربط
CD	Cardinal number	عدد اصلی
DT	Determiner	صفت اشاره

EX	Existential there	وجود
FW	Foreign word	کلمات خارجی
IN	Preposition or subordinating conjunction	حرف‌اضافه
JJ	Adjective	صفت
JJR	Adjective, comparative	صفت مقایسه
JJS	Adjective, superlative	صفت عالی
LS	List item marker	لیست نشانه ها
MD	Modal	صفت کیفی
NN	Noun, singular or mass	اسم مفرد
NNS	Noun, plural	اسم جمع
NNP	Proper noun, singular	اسم خاص مفرد
NNPS	Proper noun, plural	اسم خاص جمع
PDT	Predeterminer	از پیش تعیین کننده
POS	Possessive ending	مضاف الیه پایانی
PRP	Personal pronoun	ضمیر شخصی
PRP$	Possessive pronoun	ضمیر ملکی
RB	Adverb	قید
RBR	Adverb, comparative	قید مقایسه
RBS	Adverb, superlative	قید عالی
RP	Particle	لفظ
SYM	Symbol	سمبل
TO	to	to
UH	Interjection	حرف ندا
VB	Verb, base form	فعل فرم پایه
VBD	Verb, past tense	فعل زمان گذشته
VBG	Verb, gerund or present participle	فعل، اسم مصدر یا حال استمراری

VBN	Verb, past participle	فعل مفعولی
VBP	Verb, non-3rd person singular present	فعل غیر سوم شخص حاضر
VBZ	Verb, 3rd person singular present	فعل سوم شخص حاضر
WDT	Wh-determiner	کلمات پرسشیwh
WP	Wh-pronoun	ضمیر WH
WP$	Possessive wh-pronoun	ضمیر ملکی WH
WRB	Wh-adverb	قید WH

با ترکیب و قرار دادن انواع کلمات در کنار هم عبارات یا Phrase تشکیل داده می‌شود عبارت‌ها انواع مختلفی دارند مانند عبارت اسمی، عبارت فعلی، عبارت قیدی و ... انواع این عبارت‌ها در جدول ٤-٣ نشان داده‌شده است.

جدول .Error! No text of specified style in document-۳ انواع عبارات در جمله

نشانه	شرح انگلیسی	شرح فارسی
S	Sentence	جمله
NP	Noun Phrase	عبارت اسمی
VP	Verb Phrase	عبارت فعلی
PP	Prepositional phrase	عبارت مقدم
ADJP	Adjective phrase	عبارت وصفی
ADVP	Adverb phrase	عبارت قیدی
SBAR	Subordinate clause	شرط وابسته
SBARQ	question by Wh-element	سوالی با عناصر Wh
SQ	Yes/no questions by wh-element	سوالی بله/خیر با عناصر Wh
WHADVP	Wh-adverb phrase	عبارت قیدی با Wh

WHNP	Wh-noun phrase	عبارت اسمی با Wh
WHPP	Wh-prepositional phrase	عبارت مقدم با Wh

پس از مشخص کردن نوع کلمات بایستی نوع عبارت را به دست بیاوریم جدول ٤-٤ قوانین مربـوط به عبارت‌ها است.

جدول .Error! No text of specified style in document-٤ قوانین مربوط به عبارت‌ها

Sr. No.	Phrases	Phrases and Rules
1.	S	i. S = NP VP ii. S = NPP VP iii. S = VP iv. S = NP NPP VP v. S = NPP NPP NP VP
2.	NP	i.NP = N ii. NP = Det Adj N iii. NP = Det N iv. NP = Pron v. NP = Pron N vi. NP = Num N vii. NP = Num N N viii. NP = N Conj N ix. NP = Num N N Conj N x. NP = Det N N xi. NP = Det Adj Adj N xii. NP = Pron N N xiii. NP = Adj Pron N xiv. NP = Det Adj N N xv. NP = Det Adj N Pron xvi. NP = Neg N xvii. NP = Pron Adj N
3.	NPP	NPP = Prep NP
4.	AP	i. AP = Adj ii. AP = Adj Adj iii. AP = Adj Conj Adj
5.	APP	APP = Prep AP

6.	V	i.V = V ii. V = V V iii. V = V Adv V iv. V = V Neg V v. V = V V V V vi. V = V Conj V vii. V = V Adv viii. V = V Neg V Adv ix. V = Adv Conj Adv x. V = Adv V Neg V xi. V = V Adv Conj Adv xii. V = Adv V xiii. V = V V Adv
7.	VPP	VPP = Prep V
8.	VP	i.VP = V NP ii. VP = V VPP NP iii. VP = V NPP NP iv. VP = V NP NPP v. VP = V AP vi. VP = V NP NP VPP vii. VP = V viii. VP = V NPP ix. VP = V VPP x. VP = V NP V xi. VP = V NP VPP NP xii. VP = V VPP NPP xiii. VP = V NP NPP V NP xiv. VP = V NP AP xv. VP = V NP AP VPP xvi. VP = V NPP NPP xvii. VP = V NP V NPP xviii. VP = V VPP NP NP xix. VP = V NP NPP NPP xx. VP = V NPP NPP NPP xxi. VP = V VPP AP NPP NPP xxii. VP = V VPP NP NPP xxiii. VP = V AP NPP NPP xxiv. VP = V NP AP NPP xxv. VP = V NPP AP xxvi. VP = V VPP NP AP xxvii. VP = V AP NPP xxviii. VP = V NP VPP NP NPP xxix. VP = V NP NPP xxx. VP = V NPP VPP NP xxxi. VP = V NPP AP NPP

با استفاده از این قوانین می‌توان درخت تجزیه را رسم کرد. به‌عنوان‌مثال جمله The weather is

good را در نظر می‌گیریم ابتدا باید کلمات جمله جداشده و تگ گذاری شوند.

.| The | weather | is | good

The/DT weather/NN is/VBZ good/JJ./.

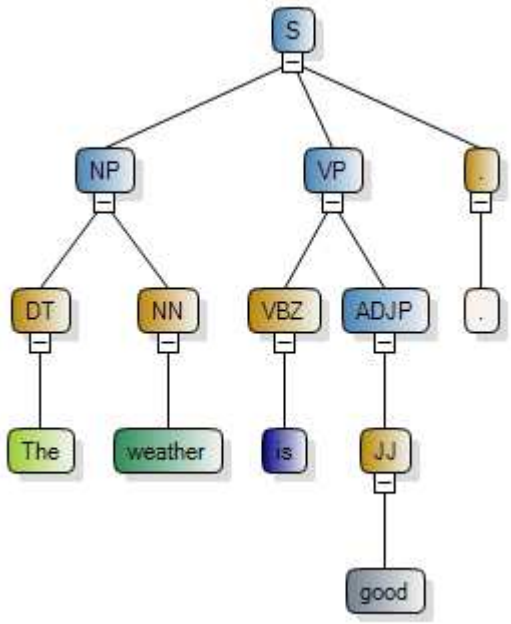
سپس با استفاده از قوانین گفته‌شده عبارت‌ها استخراج می‌شوند.

./.[NP The/DT weather/NN] [VP is/VBZ] [ADJP good/JJ]

پس از مشخص شدن عبارت‌ها برای Parse جمله از قوانین S یعنی قانون اول استفاده می‌کنیم یعنی:

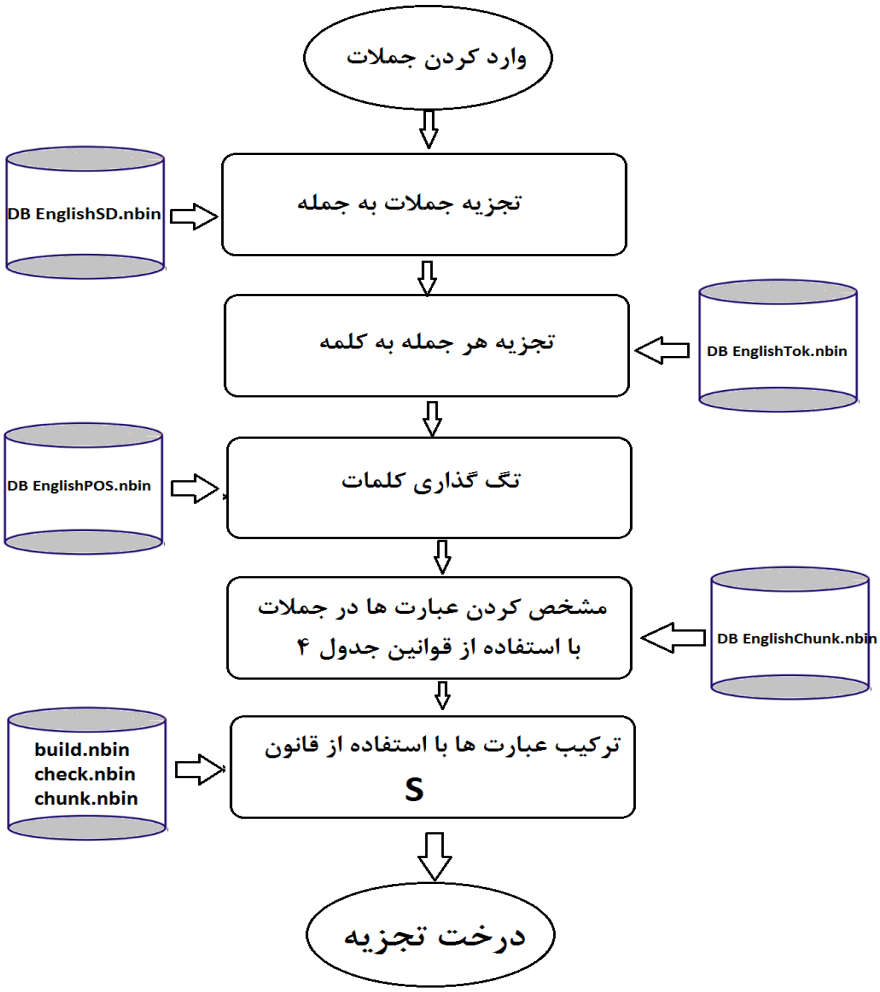
)))./)))).(TOP (S (NP (DT The) (NN weather)) (VP (VBZ is) (ADJP (JJ good

و نمودار مربوطه به شکل (٤-٢) نشان داده می‌شود.



شکل ٢-Error! No text of specified style in document. **نمودار درختی تجزیه جمله**

به‌طورکلی فلوچارت زیر را برای تعیین ساختار جمله بکار برده‌ایم.



شکل .Error! No text of specified style in document-۳ **فلوچارت روش پیاده‌سازی**