سیستم تحلیل گر متن فارسی با استفاده از شبکه عصبی SEHMM، شبکه عصبی ، مدل مارکوف مخفی

محمد امین چپ نویس (دانشگاه آزاد اسلامی واحد نجف آباد)

چکیده

مراحل عملکرد این سیستم به این ترتیب می باشد که ابتدا با عمل پیش پردازش بر روی کلمه که باعث افزایش کارایی سیستم می شود ،شروع می شود و در ادامه چگونگی آموزش شبکه عصبی برای انجام عملیات لازم خواهد بود و همچنین چگونگی استفاده از شبکه عصبی برای اعراب گذاری کلمات فارسی و در نهایت استفاده از یک نوع مارکوف مخفی (SEHMM)در تحلیل متون فارسی .

پیش پرداز

قبل از وورد کلمات به شکبه عصی ، اعمال مرحله پیش پردازش باعث افزایش کارایی سیستم می گردد .به دلیل پیچیدگی فرآینداعراب گذاری زبان فارسی ، استخراج و اعمال بعضی قوانین اعراب گذاری در افزایش بازدهی سیستم بسیار مفید است .برای مثال مجموعه حروفی وجود دارند که اگر پشت سر هم در جای مشخصی از کلمه ظاهر گردند اعراب خاصی خواهند پذیرفت .مجموعه این قوانین در یک پایگاه داده ذخیره می گردند .هر چند که این قوانین می توانند توسط شبکه عصبی استخراج گردند ، استخراج این قوانین دقت سیستم را به اندازه قابل ملاحظه ای افزایش می دهد .کلماتی در زبان عربی وجود دارند که اعراب گذاری آنها از قوانین خاصی تبعیت می کند .به علت تعداد زیاد کلمات عربی وارد شده به زبان فارسی ، جدا کردن این کلمات الزامی به نظر می رسد .به طور کلی قوانین در اعراب گذاری زبان فارسی استخراج گردید .

قوانین به کار رفته در این تحقیقات عبارتند از :

۱ حرف "ا "همیشه بیانگر واکه بلند همخوان قبلی خویش است .

۲ اگر حرف الف دو حرف مانده به انتهای کلمه در کلمه ظاهر گردد و اگر دو حرف باقیمانده همخوان باشند ، آن دو حرف ساکن خواهند بود .

۵ اگر حرف "ه "آخرین حرف یک کلمه باشد ، حرف ماقبل کلمه دارای واکه کوتاه کسره می باشد .

٦ اگر بین دو الف موجود در ی کلمه ، دو همخوان وجود داشته باشند همخوان اول دارای حرکت ساکن و همخوان

دوم دارای واکه بلند "ا "می باشد .

۷ اگر حروف "و "و یا "ی "حروف ماقبل آخر کلمه باشد ، بیانگر واکه بلند حرف ماقبل خود می باشند .

۸ اگر حروف "و "و یا "ی "حرف آخر کلمه باشند ، بیانگر واکه بلند حرف ماقبل خود می باشند .

۹ همخوان آخر هر کلمه ساکن فرض می گردد.

۱۰ همانطور که در جدول موجود در ضمیمه 1نشان داده شده است ، بعضی از همخوانهای ابتدایی کلمات وابسته به همخوان دوم کلمه با احتمال یک تنها یک نوع واکه کوتاه را خواهند پذیرفت .بنابراین هیچ ابهامی در مورد اعرابگذاری این همخوانها وجود ندارد.

۱۱ کلمات عربی بسیاری در زبان فارسی وارد شده اند .این کلمات دارای اعراب گذاری مشخصی می باشند .بنابراین کلماتی که جزو این دسته از کلمات باشند دارای الگوهای مشخصی بوده و نیازی به پردازش ندارند .قوانین استخراج شده دارای استثنائاتی می باشند که باید قبل از بکارگیری آنها بررسی گردند .این استثنائات شامل کلماتی می باشند که قوانین در مورد آنان صدق نمی کند .این کلمات در یک پایگاه داده ذخیره می گردند .

آخرین قسمت مرحله پیش پردازش شامل جداسازی پسوندها و پیشوندهای کلمات می باشد .پیشوندها و پسوندهای بسیاری در زبان فارسی وجود دارند که با اتصال آنها به کلمات ، کلماتی با

معانی متفاوت از یک کلمه ساخته می شوند .اتصال پیشوند و پسوند به کلمه تأثیری در اعراب کلمه اصلی نمی گذارد . بنابراین با جداسازی پیشوندها و پسوندها ، شبکه عصبی تنها بایستی برای کلمات اصلی آموزش ببیند و بعد از آموزش قادر خواهد بود گروه وسیعی از کلمات مشتق از یک کلمه اصلی را به درستی پاسخ دهد .

آموزش شبکه عصبی

یک فایل متشکل از متداولترین کلمات زبان فارسی برای آموزش شبکه عصبی مورد استفاده قرار می گیرد .کلمات موجود در این فایل باید به طور کامل اعراب گذاری شده باشند .قبل از آنکه شبکه عصبی شروع به آموزش کند ،هر حرف کلمات به معادل 6بیتی آن تبدیل می شود .جدول (1-5)مجموعه کامل این تبدیلها را نشان می دهد 6 .بیت معادل هر حرف به همراه 4حرف راست و 4حرف چپ مجاور وارد شبکه عصبی می شوند .برای حروفی که نزدیک دو انتهای کلمه قرار دارند به جاهای خالی یک کاراکتر خاص عمومی (Blank)وارد می شود .شبکه عصبی پردازش را برای هر کلمه مجاور به صورت مجزا انجام می دهد .بدان معنی که کلمات مجاور یک کلمه در اعراب گذاری آن کلمه دخیل نیستند .در حقیقت در این پایان نامه یک پنجره لغزان همانند آنچه در Net Talkمورد استفاده قرار گرفته ، استفاده شده است .در Net Talk نیز حرف مورد پردازش در وسط پنجره قرار می گیرد .

چگونگی پردازش کلمه "شهر "در شبکه عصبی

انتخاب 4حرف راست و 4حرف سمت چپ برای ورود به شکبه عصبی به دلیل آن است که انتخاب مقدار کمتری حروف مجاور باعث می شود که شکبه عصبی اطلاعاتی لازم را برای تصمیم گیری در اختیار نداشته باشد .به عبارت دیگر با انتخاب تعداد کمتری از حروف مجاور ممکن است در مجموعه آموزشی به ازای دو ورودی یکسان دو خروجی متفاوت وجود داشته باشد .

علاماتی در زبان فارسی وجود دارند)به طور مثال تشدید ، تنوین ، همزه ، (....و پردازشگر متن باید در صورت برخورد با چنین علاماتی واکنش مناسبی را انجام دهد .یکی از این علامات تشدید است .از این رو علامت در مواقعی استفاده می شود

که دو حرف صامت یکسان در کنار هم قرار گیرند و اولی سکون و دومی دارای یک واکه کوتاه باشد .در آن صورت حرف اول حذف شده و علامت تشدید بر روی حرف دوم قرار می گیرد .وقتی که تحلیلگر متن به این علامت برخورد می کند تشدید را حذف نموده و حرف حذف شده را جایگزین می نماید .دراین حالت حرف اول دارای علامت ساکن بوده و هیچ گونه پردازشی نیاز ندارد .

جدول (5-1)تناظر یک به یک به کار رفته بین الفبای زبان فارسی و رشته های 6بیتی

ط	010011	ا	000001
ظ	010100	ب	000010
ع	010101	پ	000011
غ	010110	ت	000100
ف	010111	ث	000101
ق	011000	ج	000110
ک	011001	چ	000111
گ	011010	ح	001000
ل	011011	خ	001001
م	011100	د	001010
ن	011101	ذ	001011
و	011110	ر	001100
ه	011111	ز	001101
ی	100000	ژ	001110
آ	100001	س	001111
	1000010	ش	010000
		ص	010001
		ض	010010

همانطور که مقدار خروجی شبکه عصبی صدای هر حرف می باشد و هر حرف می تواند دارای یک واکه بلند و واکه کوتاه باشد ، بنابراین خروجی شبکه عصبی بایستی قادر به تشخیص واکه های کوتاه و واکه های بلند یک حرف به علاوه حالتی که یک حرف دارای هیچ علامت واکه های بلند و واکه های کوتاه نباشد ، باشد .

خروجی شبکه عصبی همچنین بایستی قادر به تشخیص بعضی از استثنائات زبان فارسی نیز باشد .لیست کامل خروجیهای شبکه عصبی در جدول (5-2)آورده شده است .

جدول (5-2)خروجیهای شبکه عصبی

/e/	ـِ	001
/u/	ـُ	010
/ae/	ـَ	011

	نشانگر واکه بلند	100
	نشانگر موارد خاص	101
	سکون	110

برای تشخیص واکه بلند ، شبکه عصبی ابتدا حرف بعدی کلمه مورد پردازش را مورد بررسی قرار می دهد .اگر حرف بعدی یکی از حروف "ا "و "ی "و "بود ، حرف مرود پردازش می تواند دارای واکه بالند باشد .اما رحوف نشانگر واکه بلند گاهی به صورت همخوان نیز ظاهر می گردند .تشخیص نقش این سه حرف نیز بر عهده شبکه عصبی است .اگر سه حرف گا "و "و" "ی "بیانگر واکه بلند باشد دیگر خد به هیچگونه پردازشی توسط شبکه عصبی نیاز نداشته و شبکه عصبی بدون پردازش از روی آنها عبور می کند .خروجی شبکه عصبی باید قادر به تشخیص موارد خاص موجود در زبان فارسی نیز باشد . در زبان فارسی مواردی وجود دارد که املای کلمه با انچه کلمه تلفظ می شود تفاوت دارد به طور مثال سه حرف "خ "، "و "و "ا "اگر پشت سر هم در یک کلمه ظاهر گردند گاهی به صورت "خا "خوانده می شوند بدان معنی که حرف "و " تلفظ نمی گردد .وقتی که شبکه عصبی با چنین استثنائاتی برخورد می کند به سیستم تحلیلگر متن اعلام و سیستم تحلیلگر تغییرات لازم را در خروجی خود اعمال می کند .

بکارگیری شبکه عصبی در اعراب گذاری متنهای فارسی

هنگامی که شبکه عصبی به درستی آموزش دید ، می تواند برای اعراب گذاری متون دیگر به کار رود .در این حالت ابتدا متن به کلمات آن تقسیم شده و سپس مرحله پیش پردازش بر روی کلمات آن اعمال می شود .این پیش پردازش شامل بررسی قوانین و استثنائات آنها و نیز جداسازی پسوند و پیشوند کلمات می باشد .بعد از انجام پیش پردازش ، شبکه عصبی بر روی کلمات ، حرف به حرف ، حرکت نموده و اعراب حرفو را استخراج می نماید .بکارگیری قوانین موجود در اعراب گذاری کلمات باعث افزایش کارآیی سیستم به میزان قابل توجهی می گردد .هنگامی که شبکه عصبی شروع به اعراب گذاری کلمات می نماید ، ممکن است به دلیل وجود خطا ، کلمات اعراب گذاری شده قابل تلفظ نباشند .برای شناسایی خطاهای حاصل از پردازش شبکه عصبی ، یک مدل مارکوف پنهان 33حالته به کار گرفته شده است .

استفاده از SEHMMدر شناسایی و اصلاح خطا

یک مدل مارکوف پنهان ، مدل مارکوفی است که در آن مشاهدات دارای یک تابع توزیع براساس حالت سیستم هستند . حاصل یک مدل احتمالی 2لایه است که فرآیند زیرین قابل مشاهده نمی باشد .در مدل SEHMM، تابع مشاهده سیستم علاوه بر حالت جاری بر حالت های مجاور نیز مشروط شده است .با دانستن اطلاعات مربوط به حالتها ، تابع توزیع مشاهده سیستم مشخص می گردد.

SEHMMبرای سمبلهای گسسته مشاهده ، توسط پارامترهای زیر مشخص می گردد .در ابتدا باید تعداد حالتهای سیستم مشخص گردد .به دلیل انکه حالتها همان حروف تشکیل دهنده کلمات فارسی در نظر گرفته شده اند ، تعداد حالتهای سیستم 33حالت در نظر گرفته می شود .انتخاب 33حالت به دلیل آن است که تعداد حروف زبان فارسی 32عدد می باشد و فاصله نیز به عنوان یک حالت مجزا در نظر گرفته شده است .

فرض می شود Qمجموعه حالتهایی باشد که در تعیین تابع توزیع مشاهده سیستم در زمان iدخالت دارند .

$$Q = (q_{i-1,} q_i , q_{i+1})$$

$q_i = j1 \le j \le 33$

پارامتر دیگر تعداد سمبلهای مشاهده (V)می باشد ، که در مدل پیشنهادی 6سمبل خروجی پیشنهاد شده است 3 .عدد برای واکه های کوتاه ، 1عدد برای سکون ، 1عدد برای واکه های بلند و 1عدد برای اعلام استثنائات ، در نظر گرفته شده است . بنابراین مشاهدات سیستم به صورت زیر است :

$$V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$$

همانطور که قبلاً ذکر شد ، ارگادیک بودن مدل به معنی آن است که هر حالت از حالت دیگر با یک انتقال قابل دسترسی است .این مدل یکی از مدلهای مناسب برای شبیه سازی زبان است و راه را برای به روز سازی مدل برای کلمات جدید وراد شده به زبان هموار می سازد .

تابع توزیع احتمال مشاهدات سیستم براساس اطلاعات آماری درباره اعراب هر حرف مشروط بر حروف قبلی و بعدی تعیین می گردد) (q_t) بیانگر حالت سیستم در زمان tمی باشد. (

$$b_{lji}(k) = p(o_t = v_k | q_{t-1} = l, q_{t+1} = i)$$

$$p(o \left| Q, \lambda \right.) = \prod_{t=1}^{T} p(o_t | q_{t-1}, q_t, q_{t+1}, \lambda)$$

$$= b_{q_1 q_2 q_3 (o_2) \ldots . b_{qt-qtqo}(o_T)}$$

$$whereqo = Blank$$

همانطور که در رابطه فوق نشان داده شده است ، برای ساده سازی مدل ، مشاهدات سیستم مستقل در نظر گرفته شده اند.

در مورد سیستمهای تحلیلگر متن ، اطلاعات مربوط به تابع توزیع مشاهدات سیستم از شمارش تعداد رخداد در یک پایگاه داده حاصل می گردد .به منظور استخراج هرچه دقیقتر اطلاعات آماری ، میزان عمومیت یک کلمه و یا به بیان دیگر احتمال وقوع یک کلمه در یک متن نیز بایستی در شمارش منظور گردد .این احتمال به صورت یک وزن در شمارش تعداد رخداد در نظر گرفته می شود .

$$b_{ljo}(k) = \frac{w \in (o_t = v_k)^{\sum p_w}}{\sum_w p_w}$$

که در آن Wکلمه ای است که دارای رشته حالت p_w, Q احتمال وقوع آن است .

الگوریتمهای تخمینی که معمولاً برای آموزش مدلهای مارکوف پنهان مورد استفاده قرار می گیرد، از معیار MLبرای استخراج پارامترهای سیستم براساس مشاهدات استفاده می گردد.

$$\lambda_{ML} = \arg(\max_{\lambda} p(O|\lambda))$$

چگونگی اعراب گذاری مدل SEHMMدر شکل نشان داده شده است .

روند اعراب گذاری توسط SEHMM

این مدل برای تصحیح خطاهای شبکه عصبی بکار می رود .برای مثال اگر حرف اول یک کلمه بدون VOWELباشد ، آن کلمه را به راحتی نمی توان تلفظ نمود .برای رفع این مشکل SEHMMاعرابی برای حرف اول کلمه در نظر می گیرد . انتخاب اعراب براساس توابع توزیع مشاهدات انجام می شود .این توابع در جداولی در یک پایگاه داده ذخیره می شوند . لیست توابع توزیع مشاهده برای رشته حالت $Q = (Blank, q_t, q_{t+1})$ در ضمیمه 1نشان داده شده است .همانطور که

مشاهده می شود در مورد بعضی از رشته حالتهای Qهیچ ابهامی در مورد اعراب حرف اول وجود نداشته و تنها دارای یک حالت ممکن می باشد .با استفاده از این اطلاعات آماری می توان اعراب حرف اول کلمه را به خوبی پیش بینی نمود .
اشتباه مهم دیگری که ممکن است منجر به غیر قابل تلفظ دن کلمه گردد ، آمدن سه همخوان به دنبال هم است .اگر هیچکدام از این همخوانها دارای واکه بلند و یا واکه کوتاه نباشد کلمه حاصل غیر قابل تلفظ می گردد .در این حالت نیز SEHMMبرای یکی از همخوانها یک واکه کوتاه قرار می دهد .روند کامل اعراب گذاری در شکل نشان داده شده است .

پردازش جمله

در زبان انگلیسی تلفظ یک کلمه براساس چگونگی قرار گرفتن در جمله تغییری نمی کند .اما در زبان فارسی اعراب حرف آخر کلمات تماماً به نقش کلمات در جمله وابسته اند .حرف آخر کلماتی از نوع اسم و صفت می تواند از جمله ای به جمله دیگر تغییر اعراب دهند .تشخیص اعراب حرف آخر کلمات احتیاج به مرحله پردازش نحوی دارد که در آن کلمات با شناسایی نقش آنها در جمله اعراب گذاری می گردند .

مرجع :

آرمین غیوری ،استفاده از شبکه عصبی در تبدیل متن فارسی به گفتار و ارائه یک مدل مارکوف پنهان –دانشگاه صنعتی اصفهان