پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

مروری بر ترجمه ماشینی نورونی

نگارنده

شقایق یوسف پور

استاد راهنما

دکتر باقر باباعلی

پایان‌نامه برای دریافت درجه کارشناسی
در رشته علوم کامپیوتر

تیر ماه ۱۳۹۷

چکیده

حیات انسان‌ها در گرو برقراری ارتباط با هم نوعانشان است. از دیرباز تا کنون، انسان‌ها به واسطه برقراری ارتباط با یکدیگر توانستند بر مشکلات متعدد فایق آیند و امروزه قدرتمند ترین جانوران کره زمین شوند. انسان امروزی خواهان ارتباط هر چه بیشتر با محیط اطراف خود بوده و ابزاری در دست دارد برای غلبه بر مشکل وجود زبان‌ها ی متعدد در سرتاسر جهان و عمر کوتاه او برای یادگیری تمام آن‌ها! ترجمه، فنی ضروری برای بشریت.

هدف از ترجمه زبان‌های طبیعی به یکدیگر، فهم بهتر و بیشتر متون و گفته ها، و ایجاد ارتباط عمیق تر بین تمامی افراد سراسر جهان در طول تاریخ، با هر زبان، نژاد و فرهنگی است. اهالی این فن، همواره خدمات شایانی بر همگان ارزانی داشتند. امروزه با پیشرفت علم رایانه، گمانه زنی ها بر سر این است که کامپیوترها، مترجمان آینده باشند.

در این پروژه، مروری بر انواع روش های ترجمه ماشینی میکنیم. ابتدا روش‌های سنتی که از دیر باز مورد استفاده بودند را به اختصار توضیح می‌دهیم، سپس به تشریح روش ترجمه ماشینی نورونی؛ که انقلابی در مبحث ترجمه خودکار است، می‌پردازیم.

فهرست مطالب

فصل ۱

ضرورت ترجمه

ما در دنیایی زندگی میکنیم که شهرهای بزرگ آن به جوامع جهانی تبدیل شده اند. مردم کشورهای مختلف در یک منطقه جغرافیایی زندگی میکنند یا به جایی سفر میکنند، اما ممکن است ارتباط برقرار کردن با دیگران برایشان سخت باشد. ترجمه به آنها کمک می کند تا ارتباط آسان تری داشته باشند.

ترجمه [۱]، درک معنای یک متن و سپس تولید یک متن معادل است که همان پیام را در یک زبان دیگر تداعی میکند. در این دوره جهانی شدن که هر انسانی می‌خواهد در هر لحظه دنیای جدیدی را اکتشاف کند، اجتناب از اهمیت ترجمه غیرممکن است. ترجمه نه تنها راه پیشبرد تعاملات جهانی را ایجاد می کند، بلکه به کشورها اجازه می دهد تا روابط تعاملی را در هنگام پیشرفت در فن آوری، سیاست و غیره ایجاد کنند. با رشد اینترنت و تکنولوژی، دسترسی به مخاطبانی که هزاران مایل دورتر هستند، به سادگی امکان پذیر است. این مسئله ترجمه در زمینه‌های متنوع مانند آموزش، تجارت، ادبیات، مذهب و ... را به دنبال داشته است.

۱.۱ کاربردهای ترجمه

امروزه ترجمه کاربردهای مختلفی دارد. برای مثال:

[1] translation

- در پی رشد شرکت‌های چند ملیتی، این شرکت‌ها نیاز به رد و بدل کردن اطلاعات به سرتاسر جهان دارند. همچنین ترجمه در پر کردن شکاف ارتباطی برای رسیدن به یک مخاطب جهانی، شرکت‌ها را یاری می‌کند که با مشتریان به زبانی صحبت کنند تا بتوانند درک و با آن ارتباط برقرار کنند.
- به منظور تبادل فرهنگی : امروزه موسیقی، ادبیات، فیلم و انواع مختلف هنر و فرهنگ، با استفاده از هنر ترجمه فراتر از مرز های جغرافیایی رفته است.
- ترجمه در صنعت توریسم باعث میشود که گردشگران در کشور مقصد احساس رضایت و خوشحالی داشته باشند.
- امروزه دیپلماسی بین المللی از مهمترین جنبه‌های امور خارجی یک کشور است. ترجمه این امکان را برای تسهیل این نوع روابط بوجود می‌آورد.
- ترجمه اخبار : این‌که اخبار از نهادهای محلی، مراکز منطقه ای و حتی کشورهای مختلف باشند، ترجمه به عنوان یک ابزار موثر عمل می‌کند و بدون آن خبرها بی اثر و بی اعتبار می‌شوند.

۲.۱ چالش‌های ترجمه

ترجمه نیاز به درک عمیقی از دستور زبان و فرهنگ دارد. مترجم باید قوانین زبان و نحوه استفاده آن توسط افراد محلی را به خوبی بداند. چند مورد از شایع ترین چالش های ترجمه عبارتند از :

- ساختار زبان : هر زبان در یک ساختار تعریف شده و قوانین خاص خودش را دارد. پیچیدگی و انحصار این چارچوب با دشواری های ترجمه همراه است. برای مثال ترتیب کلمات در یک جمله انگلیسی بصورت "فاعل ـ فعل ـ مفعول" است. درحالیکه در زبان فارسی این ترتیب بصورت "فاعل ـ مفعول ـ فعل" است. در نتیجه؛ در ترجمه باید دقت شود تا به طور کارآمد ارتباط برقرار شود.

- اصطلاحات : ترجمه ی اصطلاحات سخت ترین بخش ترجمه است؛ زیرا، معنی آنها از معنی ظاهری کلمهها بدست نمی آید.
- کلمات مرکب : این کلمات از ترکیب دو یا چند کلمه ایجاد می‌شوند که در بعضی از مواقع معنی آن به معنی تک تک کلمات تشکیل دهنده آن، ربطی ندارد.
- نام های نا موجود : برخی از زبانها ممکن است شامل نام ها و یا اصطلاحاتی باشند که مفهوم شان در زبانهای دیگر وجود نداشته باشد.
- افعال دوـ کلمه ای : در زبانهایی مانند انگلیسی، افعال دو حرفی وجود دارند که معنی آنها برابر معنی هر کدام آنها بصورت جدا نیست.
- معنی چندگانه : برخی کلمات در تمامی زبانها وجود دارند که چند معنی مختلف دارند. مفهوم آنها بصورت جداگانه استخراج نمی‌شود و باید در جمله و با توجه به معنی سایر اجزای جمله، به آن پرداخت.
- کنایه : کنایه ها اگر کلمه به کلمه ترجمه شوند، معنی مورد نظر خود را از دست می‌دهند. پس باید به گونه ای ترجمه شوند که معنی اصلی آنها حفظ شود.

فصل ۲

آشنایی با ترجمه ماشینی

در این فصل به مروری بر روش‌های ترجمه ماشینی می پردازیم و سپس روش ترجمه ماشینی نورونی را به تفصیل شرح می‌دهیم.

۱.۲ تاریخچه

انقلاب اطلاعاتی، نوآوری‌های تکنولوژی و توسعه صنایع زبان، توسعه چند زبانی را بوجود آورده است. استفاده از ترجمه ماشین با کمک بسیاری از تکنیک‌های متنوع و جدید، با پیشرفت بی سابقه ای همراه بوده است. با این حال هدف اصلی محققان در محیطهای تحت سلطه ی اینترنت، توسعه سریع سیستم های ترجمه ای است که دقیق و موثر باشند.

تاریخچه مدرن ترجمه خودکار به اوایل دوران پس از جنگ جهانی دوم بر می‌گردد. زمانی که تلاش برای یافتن راهی سریع برای ترجمه اطلاعات؛ افزایش یافت. در سال ۱۹۴۷ دانشمند و ریاضی دان آمریکایی وارن ویور[1]، نامه ای به سایر همکارانش درباره ی توانایی کامپیوتر برای تبدیل یک زبان به زبانی دیگر با به کار گیری منطق، رمز نگاری، ترکیبیات و الگوریتم ها ی زبانی نوشت.

[1] Warren Weaver

در سال ۱۹۵۰ یک تیم تحقیقاتی از دانشگاه جورج تاون[۲] و آی بی ام[۳] شکل گرفت. در سال ۱۹۵۴ آن ها توانستند چند ده عبارت روسی را به انگلیسی توسط ماشین ترجمه کنند. در دهه ۷۰ و ۸۰ میلادی؛ با به کار گیری ابزار جدید، دانشمندان توانستند فرآیند ترجمه ماشینی را تسهیل کنند. در دهه ۹۰ فراگیر شدن اینترنت و دسترسی همگان به کامپیوترهای قدرتمند ارزان موجب پیشرفت ترجمه ماشینی شد.

۲.۲ روش‌های ترجمه ماشینی

ترجمه ماشینی[۴] نرم افزاری است که به طور خودکار متن را از یک زبان به زبان دیگری ترجمه می‌کند. انواع مختلفی از ترجمه ماشینی وجود دارد که معروف ترین آن‌ها، ترجمه ماشینی قاعده مند و ترجمه ماشینی بر اساس متن است.

۱.۲.۲ ترجمه ماشینی قاعده مند

[۵]

این روش ترجمه، اولین روش در حوزه ی ترجمه ماشینی است. یک سیستم ترجمه ماشین قاعده مند شامل مجموعه ای از قواعد، کلمات و برنامه نرم افزاری برای پردازش قواعد است. قواعد زبان توسط زبان شناسان نوشته می‌شوند و نقش به سزایی در مراحل مختلف ترجمه؛ مانند پردازش نحوی و تفسیر معنایی دارند.

سه نوع سیستم ترجمه قاعده مند داریم :

۱. ترجمه ماشینی مبتنی بر لغت نامه[۶] : جمله ورودی را طبق قوانین پایه ای به جمله خروجی در زبان دیگر می‌نگارد

[2] GeorgeTown
[3] International Business Machine
[4] Machine Translation
[5] Ruled-Based Machine Translation
[6] Dictionary-Based Machine Translation

۲. ترجمه ماشینی مبتنی بر انتقال[۷] : تجزیه و تحلیل نحوی و مورفولوژیکی را به کار می‌گیرد.

۳. میان زبانی[۸] : معنی انتزاعی جمله و مفهوم را در نظر می‌گیرد.

ساختار

برای نمایش ساختار جمله از درخت استفاده می‌شود. یک جمله انگلیسی از دو قسمت گروه اسمی[۹] و گروه فعلی[۱۰] تشکیل می‌شود. قوانین بازنویسی توصیف می‌کنند چه ساختار درختی برای یک جمله مجاز است؛ زیرا، تنها جمله های با ساختار درست، صحیح ترجمه خواهند شد.

قواعد زبانی به صورت قواعد مستقل از متن بیان می‌شوند و از فرم های منطقی در تفسیر معنایی استفاده می‌شود.

نحوه ی ترجمه

ترجمه در روش قاعده مند، توسط تطبیق الگو[۱۱] ی قوانین انجام می‌شود. موفقیت زمانی به دست می آید که از قوانین بی‌حاصل جلوگیری کنیم. برای فهم زبان به دانش و استدلال نیاز است؛ دانش عمومی باعث می‌شود بتوانیم مشکلات تفسیری مانند ابهام معنایی را حل کنیم. دانش محتوایی می‌تواند برای تعیین مرجع عبارات اسمی و معنی درست کلمات بر اساس وضعیت فعل، استفاده شود. روش اصلی روش ترجمه ماشینی قاعده مند، مبتنی بر ارتباط ساختار جمله ورودی با ساختار جمله خروجی با حفظ معنای آنها است. برای مثال؛ اگر قرار باشد جمله ای در زبان انگلیسی را به جمله ای در زبان فرانسه ترجمه کنیم، ابتدا باید هر کلمه را به معادل فرانسوی اش بنگاریم. سپس با در نظر گرفتن قوانین نحوی زبان انگلیسی و فرانسه، ترجمه را انجامی دهیم. برای مثال جمله مقابل را در نظر بگیرید :

I go to school by bus.

[7] Transfer-Based Machine Translation
[8] Interlingua
[9] Noun Phrase
[10] Verb Phrase
[11] Pattern Matching

در مرحله اول، نقش کلمات در جمله را مشخص می‌کنیم:

، go = v. ، to = prep. ، school = n. ، by = prep. ، bus = n.
I = pron.

سپس نقش نحوی آن‌ها را می‌یابیم :

go = simple present ، I = pronoun of first person singular
... ، singular person

با parse کردن می‌توانیم نگاشت نحوی جمله مبدا و مقصد را انجام دهیم.
در مرحله بعد، هر کلمه به معادل فرانسوی اش نگاشته می‌شود :

I = Je ، to = à ، school = l'école ، by = en

سپس کلمات را به ترتیب مناسب در جمله قرار می‌دهیم :

Je vais à l'école en bus.

مزایا و معایب

از مزایای این روش می‌توان به موارد زیر اشاره کرد :

- زبان‌هایی که هیچ گونه متون مشترکی ندارند را می‌توان ترجمه کرد.
- ترجمه در این روش، مستقل از حوزه است؛ زیرا، قواعد زبان به طور کلی نوشته می‌شوند و مخصوص حوزه معنایی یا نحوی خاصی نیستند.
- از آنجائیکه قوانین بصورت دستی نوشته می‌شوند؛ برای عیب یابی و اشکال زدایی، به راحتی می‌توان با تغییر قوانین متوجه اشکال شد.
- محدودیتی برای رفع خطاها وجود ندارد؛ زیرا، هر خطایی که بروز کند، با تصحیح قانون مربوط به آن یا حتی نوشتن قانون جدید می‌توانیم آن را رفع کنیم.
- قابلیت استفاده دوباره : از آنجا که سیستم‌های ترجمه قاعده مند، به طور کلی از تجزیه و تحلیل زبان مبدا و انتقال آن به زبان مقصد استفاده می‌کنند، تجزیه و تحلیل زبان مبدا و بخش‌های تولید زبان مقصد می‌تواند بین سیستم‌های ترجمه مشترک باشد.

از معایب این روش نیز می توان از :

- ساختن فرهنگ لغت هزینه بر است.
- قسمتی از اطلاعات زبانی باید دستی تنظیم شود.
- در سیستم‌های بزرگ، برخورد با قوانین و تعاملات آن‌ها دشوار است.

نام برد.

۲.۲.۲ ترجمه ماشینی مبتنی بر متن

[۱۲]

کرپس[۱۳] یک پایگاه داده در مقیاس بزرگ است که شامل حجم بزرگی از اطلاعات زبانی است و برای بازیابی توسط کامپیوتر استفاده می‌شود. ترجمه ماشینی مبتنی بر متن تلاش می‌کند تا نقص سیستم‌های ترجمه ماشینی سنتی را کاهش داده و کارایی و دقت آن‌ها را بهبود بخشد. ترجمه ماشینی مبتنی بر متن به دو روش تقسیم می‌شود :

ترجمه ماشینی مبتنی بر مثال

[۱۴]

در این روش از پایگاه داده ای استفاده می‌شود که شامل لغات و جملاتی است که از قبل ترجمه شده اند. یه طور کلی ترجمه ماشینی مبتنی بر مثال از سه بخش تشکیل‌می شود :

۱. تطبیق ماژول[۱۵] : در این بخش، فرهنگ لغات شامل مثال‌های از پیش ترجمه شده جستجو می‌شود تا متون شبیه به جمله ورودی در آن پیدا شود.

۲. باز ترکیبی[۱۶] : متون مشابه بازیابی شده از پایگاه داده، با هم ترکیب می‌شوند تا یک جمله را بسازند.

[12] Corpus-Based Machine Translation
[13] corpus
[14] Example-Based Machine Translation
[15] module matching
[16] recombination

۳. هموار سازی[۱۷] : در راستای این که جملات بدست آمده کاملا از نظر دستوری مطابق زبان مقصد باشد و خطاها کاهش یابد، مجموعه ای از پردازش‌ها بر روی جمله بدست آمده انجام می‌شود.

ترجمه ماشینی آماری

[۱۸]

مدل اولیه ی ترجمه ماشینی آماری بر اساس تئوری بیز[۱۹] بوده است که این دیدگاه را مطرح می‌کند که هر جمله در یک زبان ممکن است ترجمه جمله ی دیگری در زبان دیگر باشد و مناسب‌ترین ترجمه آن است که بیش‌ترین احتمال را به دست آورده باشد.

ایده این روش از نظریه اطلاعات می آید. یک مستند[۲۰] با توجه به توزیع احتمالاتی $p(e|f)$ که f جمله مورد نظر در زبان مبدا و e ترجمه ی f است، ترجمه می‌شود.

$$p(e|f) = p(f|e)p(e)$$

$p(f|e)$ احتمال این که جمله مقصد ترجمه ی جمله مبدا باشد

$p(e)$ احتمال حضور جمله مقصد در زبان

حال، بهترین ترجمه $\tilde{a}$ با انتخاب ترجمه‌هایی که بالاترین احتمال را دارند، بدست می‌آید.

$$\tilde{a} = argmaxp(e|f), e \in \hat{a}$$

$\hat{a}$ تمام جملات در زبان هستند

ترجمه ماشینی آماری به سه نوع مختلف تقسیم می‌شود :

۱. ترجمه ماشینی مبتنی بر کلمه[۲۱] :

در این روش، واحد اساسی ترجمه؛ کلمه است. معمولا تعداد کلمات جمله‌های مختلف، متفاوت است. نسبت طول توالی کلمات ترجمه

[17] smoothing
[18] Statistical Machine Translation
[19] Bayes Theory
[20] document
[21] Word-Based Machine Translation

شده، باروری[۲۲] نامیده می‌شود که بیان می‌کند هر کلمه بومی؛ چند کلمه خارجی تولید می‌کند. به طور کلی، نظریه اطلاعات[۲۳] بیان می‌کند که تمامی این کلمات یک مفهوم مشترک دارند. در عمل، این مسئله درست نیست. برای مثال؛ کلمه "شیر" در فارسی می‌تواند کلمات "lion" ، "milk" و "faucet" را تولید کند؛ ولی، به کار بردن تمامی این کلمات در یک مکان درست نیست و باید با توجه به معنی مورد نظر، کلمه درست به کار برده شود.

روش مبتنی بر کلمه، نمی تواند زبان‌هایی با باروری متفاوت را به یکدیگر ترجمه کند. سیستم‌های مبتنی بر کلمه می‌توانند به سادگی برای مقابله با باروری بالا ساخته شوند. همچنین؛ می توانند یک کلمه را به چند کلمه بنگارند و نه بر عکس. برای مثال؛ یک کلمه در زبان انگلیسی ممکن است چندین هم‌معنی در زبان فارسی داشته باشد، یا حتی هیچ معادل فارسی نداشته باشد. اما هیچ راهی نیست که بتوانیم دو کلمه انگلیسی را به هم بچسبانیم تا هم‌معنی یک کلمه فارسی بشود.

۲. ترجمه ماشینی مبتنی بر عبارت[۲۴] :

در این روش؛ هدف، کاهش محدودیت‌های مدل مبتنی بر کلمه است. به صورتی که تمامی توالی‌های کلمات با طول‌های مختلف، یک‌جا ترجمه شوند. توالی کلمات، عبارت نامیده می‌شوند. اما معمولا عبارات زبان شناسی نیستند. عباراتی هستند که با استفاده از روش‌های آماری از منابع متون یافت می‌شوند؛ زیرا، نشان داده شده است که محدود کردن عبارات به عبارات زبان شناسی، کیفیت ترجمه را کاهش می‌دهد.

۳. ترجمه ماشینی مبتنی بر نحو[۲۵] :

این روش؛ بر مبنای ترجمه ی واحدهای نحوی، به‌جای کلمات یا رشته‌های کلمات است. ایده ی این روش در ترجمه ماشینی بسیار قدیمی است ، اگر چه عدد آماری آن تا زمان ظهور تجزیه کننده‌های تصادفی[۲۶] قوی در دهه ۲۰ میلادی ، افزایش داشته است.

[22] fertility
[23] Information Theory
[24] Phrase-Based Machine Translation
[25] Syntax-Based Machine Translation
[26] stochastic parser

مزایا و معایب

- ترجمه‌های روانتر و رساتر
- کارامدی در استفاده از منابع متنی موجود

از مزایای این روش هستند.
همچنین از معایب نیز می‌توانیم به نکات زیر اشاره کنیم :

- ساختن منبع متنی ممکن است هزینه بر باشد.
- یافتن و رفع برخی از خطا ها دشوار است.
- این روش برای ترجمه زبان‌هایی که ترتیب کلمات شان با هم متفاوت است، خیلی کارا نیست.

۳.۲.۲ ترجمه ماشینی نورونی

[۲۷]

ترجمه ماشینی نورونی یکی از رویکردهای ترجمه ماشینی آماری است که با استفاده از یک شبکه عصبی مصنوعی بزرگ، احتمال دنباله ای از کلمات را پیش بینی می‌کند.
در فصل بعد، مفصل به این شیوه از ترجمه ماشینی خواهیم پرداخت.

[27] Neural Machine Translation

فصل ۳

ترجمه ماشینی نورونی

یک پیشرفت عمده اخیر در ترجمه ماشینی آماری، ترجمه به وسیله ی شبکه عصبی است.

در این فصل تعدادی از تکنیک‌های مدل سازی شبکه‌های عصبی را معرفی کرده، و نحوه ی اعمال آن‌ها به مشکلات ترجمه ماشینی را توضیح می‌دهیم.

۱.۳ تاریخچه

در طی تحقیقات در حوزه شبکه‌های عصبی در دهه‌های ۱۹۸۰ و ۱۹۹۰ ، ترجمه ماشینی همواره مورد توجه محققان؛ برای یافتن روش‌های نو، بود. در حقیقت، روش ارائه شده توسط فورکادا[1]، نکو[2] [۸] و کاستانو[3] [۹] ، شبیه رویکردهای ترجمه ماشینی نورونی فعلی بودند اما هیچ کدام از این مدل‌ها روی داده‌های بسیار بزرگ آزمایش نشده بودند تا نتایج قابل قبول تولید کنند. همچنین پیچیدگی محاسباتی که به مراتب بیش از منابع محاسباتی آن دوران بود، باعث شد این کار برای دو دهه رها شود.

در طول این مدت، رویکردهای داده ای مانند ترجمه ماشینی آماری مبتنی بر کلمه، ترجمه ماشین را به ابزاری مفید برای بسیاری از برنامه‌های کاربردی

[1] Forcada
[2] Ñeco
[3] Castaño

تبدیل کرد. بازگشت دوباره ی روش‌های نورونی در ترجمه ماشینی با یکپارچگی مدل‌های زبانی نورونی و سیستم‌های سنتی ترجمه ماشینی آماری آغاز شد. اگرچه، ایده‌های مطرح شده در این دوران، به دلیل محدودیت‌های محاسباتی به کندی مورد پذیرش قرار می‌گرفتند.

نخستین قدم ها در جهت افزایش استفاده از روش‌های نورونی و ترک کامل روش‌های آماری، استفاده از مدل‌های پیچشی[۴] (کانولوشن) و مدل‌های توالی به توالی[۵] بود. این روش‌ها ترجمه معقولی برای جملات کوتاه می‌دادند اما با افزایش طول جملات، درستی جواب‌ها کاهش می‌یافت. با اصلاحات جدید مانند روش‌های کدگذاری جفت بایت[۶]، ترجمه ماشینی نورونی به مراحل تازه ای رسید.

در طی یک تا دو سال، تمامی تحقیقات در حوزه ترجمه ماشینی، به سمت روش‌های عصبی رفت. در کنفرانس ترجمه ماشین در سال ۲۰۱۵، تنها یک سیستم ترجمه نورونی ارائه شد زیرا از سیستم‌های سنتی آماری اشباع شده بود. یک سال بعد؛ در ۲۰۱۶، یک سیستم ترجمه نورونی تقریبا در تمامی ترجمه‌های دو به دو ی زبان‌ها، برنده رقابت‌ها شد. در سال ۲۰۱۷، تمامی شرکت کننده های این کنفرانس، سیستم‌های ترجمه نورونی بودند.

در حال حاضر، پژوهش‌ها در حوزه ترجمه نورونی ادامه دارد و در آینده نیز تحقیق درباره موارد بیشتری از جمله مدل‌های عمیق تر برای یاد گیری ماشین[۷] و مدل‌های آگاهانه تر نسبت به مسائل زبان شناختی، ادامه خواهد داشت.

طیف گسترده ای از ابزارهای موجود برای تحقیق، توسعه، و استقرار سیستم‌های ترجمه ماشینی نورونی وجود دارد. برخی از این ابزار ها عبارتند از :

- Nematus [۴]
- OpenNMT [۵]
- Tensor to Tensor

[4] Convolutional Neural Network
[5] sequence-to-sequence models
[6] Byte Pair Encoding
[7] Machine Learning

۲.۳ مروری بر شبکه ها ی عصبی

۸

شبکه عصبی یک روش یاد گیری ماشین است که مقداری ورودی گرفته و خروجی را پیش بینی می‌کند.
در این بخش به توضیح بسیار مختصری درباره این شبکه ها می پردازیم.

۱.۲.۳ مدل خطی

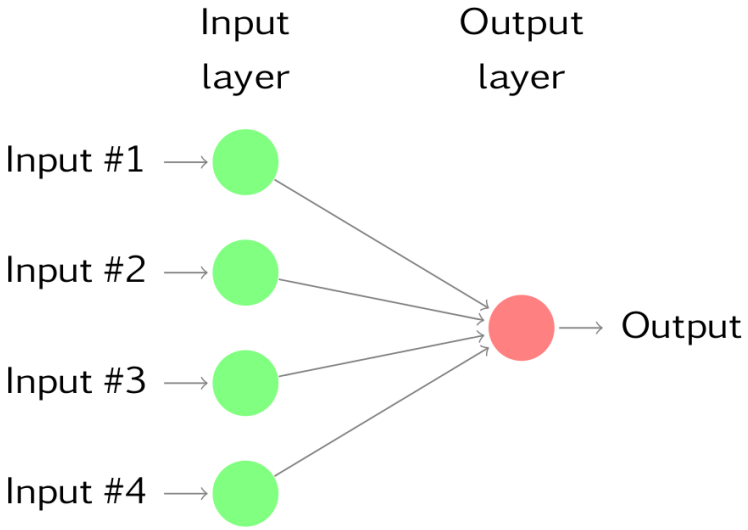
۹

مدل خطی، عنصر اصلی ترجمه ماشینی آماری است. یک ترجمه بالقوه x از یک جمله، بصورت مجموعه ای از ویژگی‌های $h_i(x)$ نشان داده می‌شود. هر ویژگی توسط پارامتر λ_i وزن می‌شود تا امتیاز بگیرد. با چشم پوشی از تابع نمایی که برای تبدیل مدل خطی به مدل خطی لگاریتمی استفاده می‌شود، فرمول این مدل به صورت زیر است :

$$score(\lambda, x) = \sum_j \lambda_j h_j(x)$$

به صورت گرافیکی، یک مدل خطی می‌تواند به صورت شکل ۱.۳ نشان داده شود. گره های ورودی؛ مقدار ویژگی‌ها، یکان‌ها؛ وزن‌ها و امتیاز نهایی همان گره خروجی است.[۳]
از مدل‌های خطی برای ترکیب قسمت های مختلف یک سیستم ترجمه ماشین مانند مدل زبانی، مدل ترجمه عبارات و یا ویژگی‌هایی مانند طول جملات و... استفاده می‌شود. متدهای آموزش[۱۰]، یک مقدار وزنی λ_i به هر ویژگی $h_i(x)$ ، با توجه به اهمیت آن‌ها نسبت می‌دهند به طوری که هر چه امتیاز

[8] Neural Network
[9] Linear Model
[10] training



شکل ۱.۳: شمایی از یک شبکه عصبی خطی

بیشتر باشد ترجمه بهتری داریم. در ترجمه ماشینی آماری به این روش میزان سازی[۱۱] گویند.[۳]

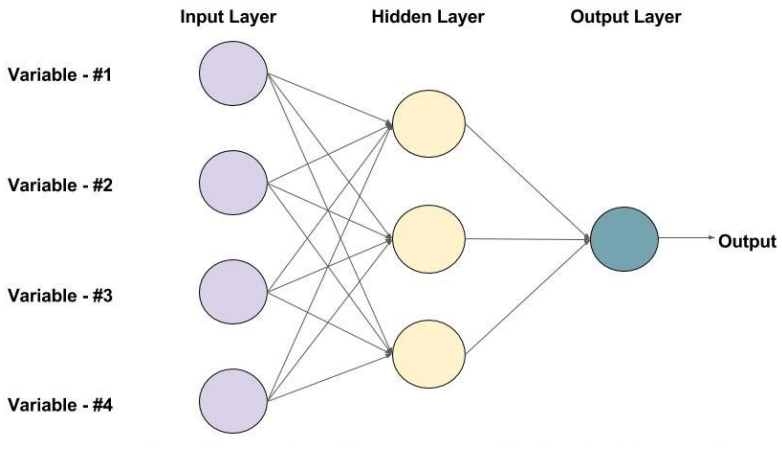
مدل‌های خطی به ما اجازه نمی‌دهند که روابط پیچیده تری بین ویژگی‌ها تعریف کنیم. فرض کنیم برای جملات کوتاه؛ مدل زبانی نسبت به مدل ترجمه، اهمیت کمتری دارد. یا میانگین احتمالات ترجمه عبارات، بالاتر از ۰.۱ باشد معقول، و کمتر از آن بسیار بد است. نخستین مثال وابستگی بین ویژگی‌ها و مثال دوم، رابطه غیر خطی ویژگی‌ها و تاثیرش بر امتیاز نهایی را نشان می‌دهد. مدل‌های خطی نمی‌توانند این مسائل را حل کنند.

۲.۲.۳ مدل چند لایه ای

[۱۲]

در این مدل از شبکه‌های عصبی به جای این که مقادیر خروجی مستقیما از ورودی به دست آیند، از یک لایه مخفی[۱۳] استفاده می‌شود. آن را مخفی

[11] tunning
[12] Multiple Layers
[13] hidden layer



شکل ۲.۳: شمایی از یک شبکه عصبی با لایه مخفی

گویند زیرا در دادههای آموزشی ، ما تنها قادر به دیدن ورودی و خروجی هستیم و نه مکانیزمی که آنها را به هم مرتبط میکند.

طبق شکل ۲.۳ ، ابتدا یک ترکیب خطی از ورودیهای وزن دار برای محاسبه هر کدام از گرههای مخفی محاسبه میشود. سپس یک ترکیب خطی از گرههای مخفی برای تولید خروجی محاسبه میشود.

۳.۳ مدلهای زبانی نورونی

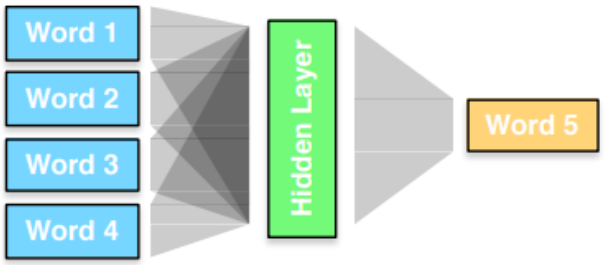
[۱۴]

شبکه عصبی یک روش بسیار قدرتمند برای مدل سازی توزیع احتمالی شرطی با ورودیهای چندگانه است$p(a|b.c.d)$. در شبکه عصبی، نقاطی در دادههای آموزشی که مشاهده نمیشوند، اشکالی ایجاد نمیکنند. اما روشهای سنتی آماری، ممکن است برای این نوع داده هایی که تعدادی از آنها وجود ندارند، خوشه بندی[۱۵] را برای رفع مشکل استفاده کنند که خوشه بندی نیز نیاز به احاطه کامل بر مسئه دارد. [۳]

مدلهای n ـ گرام ، احتمال یک کلمه را به ضرب احتمالات آن کلمه در

[14] Neural Language Models

[15] clustering



شکل ۳.۳: نمایی از یک مدل زبانی نورونی : کلمه جدید بر اساس کلمات قبلی پیش بینی می‌شوند.[۳]

متون قبلی کاهش می‌دهد $p(wi|w_{i-4}, w_{i3}, w_{i2}, w_{i1})$. چنین مدل‌هایی یک مثال اصلی برای توزیع احتمالی مشروط است که ما اغلب نقاط داده ای نداریم و خواهان خوشه بندی اطلاعات هستیم.

در مدل‌های زبانی آماری، سعی بر این است که با تخمین‌های نزولی از مدل‌های مرتبه بالاتر برای متعادل کردن محاسبات غلط ناشی از محاسبات مرتبه پایین تر؛ مانند ۲ ـ گرام ، استفاده شود. حال، از شبکه‌های عصبی برای کمک استفاده می‌کنیم.

۱.۳.۳ مدل زبانی پیش‌ـخور

[۱۶]

در شکل ۱.۳ شمایی از یک مدل زبانی نورونی ۵ ـ گرام نشان داده می شود. گرهای شبکه که اطلاعات متنی دارند به یک لایه مخفی متصل اند که این لایه خود به لایه خروجی ای که کلمه ترجمه شده را پیش بینی می‌کند، متصل است.

نمایش کلمات

حال سوال اصلی این است که کلمات را چطور باید نمایش دهیم؟ گرها در شبکه‌های عصبی، مقادیر عددی دارند، در حالی که کلمات مجموعه گسسته

[16] Feed-Forward Neural Language Model

ای از یک مجموعه واژگان بسیار بزرگ اند. برای نمایش کلمات، از کدهای شناسایی[17] نمی‌توانیم استفاده کنیم؛ زیرا، یک شبکه عصبی توکن ۱۲۴۳۲۱ را مشابه توکن ۱۲۴۳۲۲ تلقی می‌کند، در صورتی که این اعداد کاملا تصادفی اند. این مسئله برای کدگذاری بیتی نیز برقرار است. ممکن است دو کد بیتی بسیار شبیه به هم باشند ولی در عمل هیچ ربطی به هم نداشته باشند.

روشی دیگر این است که ما کلمات را با بردارهایی با ابعاد بزرگ نشان دهیم. به طوری که هر کلمه یک بعد باشد و اگر آن کلمه با ان بعد همخوانی داشت، ۱ و در غیر این صورت ۰ می‌گذاریم . به این بردار، بردار تک عنصری گویند.

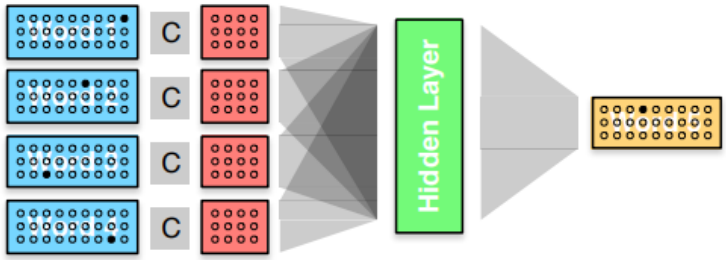
این بردار ها بسیار بزرگ اند و عموما خالی. می توانیم کلمات را به واژگان متداول، محدود کنیم و سایر کلمات را در توکنی دیگر نگهداری کنیم. همچنین می‌توانیم از گروه بندی کلمات برای کاهش ابعاد بردار ها استفاده کنیم.

اشتراک معنایی بین کلمات، یک لایه دیگر بین لایه ورودی و لایه مخفی معرفی می‌کند. در این لایه هر کلمه به طور جداگانه بر یک فضای با ابعاد کمتر، نگاشت می‌شود. برای تمامی کلمات محتوایی، یک ماتریس وزن یکسان استفاده می‌شود. در نتیجه، یک نمایش برای هر کلمه؛ مستقل از جایگاهش در جمله، به دست می‌آید. به این نحوه نمایش، تعبیه کلمه[18] گویند. [۳]

کلماتی که در محتوای مشابه هستند، تعبیه کلمه مشابه هم دارند. برای مثال، اگر داده آموزشی برای یک مدل زبانی که به طور مکرر دارای n ـ گرام‌های زیر باشد:

- the dog jumped
- the cat jumped
- child hugged the cat
- child hugged the dog

[17] token ID
[18] word embedding



شکل ۴.۳: نمونه ای از معماری مدل زبانی شبکه عصبی پیش_خور : کلمات محتوایی $(w_{i-1}, w_{i-2}, w_{i-3}, w_{i-4})$ در قالب بردارهای تک عنصری نشان داده می شوند، و سپس در یک فضای پیوسته توسط تعبیه کلمه، نگاشت می‌شوند. کلمه پیش بینی شده که توسط لایه مخفی محاسبه می‌شود، بصورت بردار تک عنصری است.[۳]

در این حالت، مدل زبانی می تواند از این مسئله که سگ و گربه در محتوای مشابه آمده اند، منتفع شود و از این رو تا حدی قابل تعویض هستند. اگر بخواهیم از محتوای شامل "سگ" برای پیش بینی استفاده کنیم، ولی در محتوا "گربه" داشته باشیم، باز هم از این مسئله استقبال می‌کنیم . تعبیه کلمه باعث تعمیم کلمات و در نتیجه پیش بینی‌های قوی حتی در محتوای از قبل دیده نشده، می‌شود.

معماری شبکه عصبی

تصویر ۲.۳ شمایی از یک معماری تکامل یافته از مدل پیش_خور است که شامل کلمات محتوایی (به صورت بردارهای تک عنصری)، لایه ی تعبیه کلمه ، لایه مخفی و لایه پیش بینی خروجی می‌شود.

کلمات محتوایی؛ در ابتدا، به صورت بردارهای تک عنصری کد گذاری می‌شوند. سپس به صورت ورودی به یک ماتریس C داده می‌شوند که حاصل آن برداری از اعداد اعشاری است که به آن تعبیه کلمه گویند. این بردار تعبیه شده، به طور معمول بین ۵۰۰ تا ۱۰۰۰ تا گره دارد.

به دلیل اینکه بردارهایی که در ماتریس وزن C ضرب می‌شوند، بردارهای تک عنصری هستند، بیشتر ورودی‌ها صفر است. در نتیجه؛ عملا، ستونهایی

از ماتریس را انتخاب می‌کنیم که با کد شناسایی کلمه ورودی مطابقت داشته باشد. پس در اینجا نیازی به تابع فعالسازی نیست.

ماتریس وزن $C(w_j)$ برای تعبیه کلمه از طریق کد شناسایی کلمات (w_j)، اندیس گذاری می شوند.

$$C(w_j) = Cw_j$$

در مدل پیش‌ـ‌خور، برای نگاشت به لایه مخفی، تمامی تعبیه کلمه ها $C(w_j)$ ی ورودی باید به هم بچسبند. برای این کار از تابع فعالسازی tanh استفاده می‌کنیم.

$$h = \tanh(b_h + \sum_j H_j C(w_j))$$

لایه خروجی به صورت یک احتمال توزیع شده بر روی کلمات کار می‌کند. همانند قبل، برای هر گره i و مقادیر مخفی h_i و وزن w_{ij}، ترکیب خطی s_i به صورت $S = Wh$ محاسبه می‌شود. حال برای اطمینان از این که این احتمال توزیعی خوب است و تمامی مقادیر حداکثر مقدار ۱ دارند، از تابع فعالسازی softmax استفاده می‌کنیم.

$$p_i = softmax(s_i, \vec{s}) = \frac{e^{s_i}}{\sum_j e^{s_i}}$$

مدل فوق بسیار شبیه مدل زبانی بنگیو[۱۹] است. در این مدل نیز پیوندی بین کلمه محتوایی و کلمه خروجی اضافه می‌شود. در نتیجه رابطه $s = Wh$ به صورت زیر تغییر می‌کند:

$$s = Wh + \sum_j UC(w_j)$$

داشتن این چنین پیوندهای مستقیمی، سرعت آموزش را افزایش می‌دهد اما کارایی را بهبود نمی‌بخشد. [۳]

آموزش

یک مدل زبانی نورونی با پردازش تمامی n ـ گرامها در داده آموزشی، آموزش داده می‌شود. برای هر n ـ گرام کلمات محتوایی را در شبکه می‌فرستیم و

[19] Bngio et al. (2003)

خروجی شبکه را با بردار تک عنصری کلمه صحیح مطابقت می‌دهیم. سپس وزن‌ها بوسیله ی انتشار به عقب[20] آپدیت می‌شوند.

مدل‌های زبانی عموما با پیچیدگی سنجیده می‌شوند که مرتبط است با احتمالی که به یک متن داده می‌شود. در نتیجه، هدف از آموزش یک مدل زبانی، افزایش احتمال[21] داده‌های آموزشی است.

در حین آموزش یک محتوای $x(w_{i-1}, w_{i-2}, w_{i-3}, w_{i-4})$، مقدار درست بردار تک عنصری $\vec{y}$ را داریم. برای هر $(x, \vec{y})$ احتمال آن را داریم :

$$L(x, \vec{y}; W) = -\sum_k y_k log p_k$$

و چون فقط یکی از مقادیر y_k برابر ۱ است و بقیه ۰ اند، حاصل برای کلمه درست ،k احتمال p_k می‌شود. با تعریف احتمال به صورت فوق، وزن‌ها آپدیت می‌شوند حتی وزن آن‌هایی که کلمه خروجی را اشتباه پیش بینی می‌کنند. [۳]

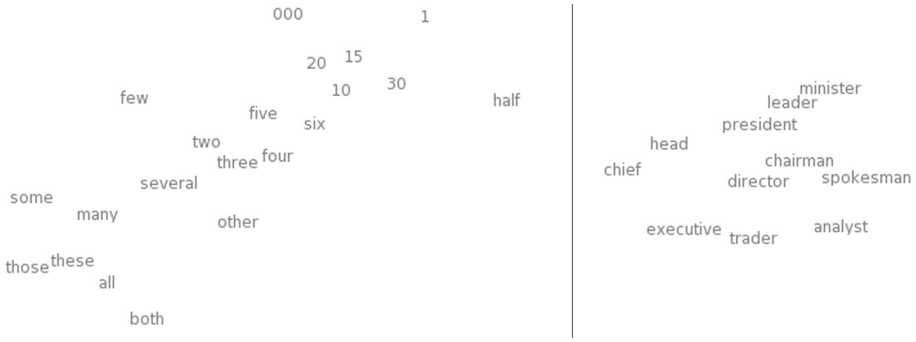
۲.۳.۳ تعبیه کلمه

نقش تعبیه کلمه در ترجمه ماشینی نورونی و بسیاری از پردازش‌های زبان طبیعی بسیار با اهمیت است. در اینجا، آن‌ها را به این گونه مطرح می‌کنیم که کلمات در فضای با ابعاد بالا؛ ۵۰۰ تا ۱۰۰۰ بعد، محاسبه شده اند.

نقش تعبیه کلمه‌ها را در نظر بگیرید. آن‌ها کلمات محتوایی را نمایش می‌دهند که مارا قادر می‌سازند کلمات را در محتوایی که در آن قرار گرفتند، پیش بینی کنیم. مثال سگ و گربه را به یاد آورید. چون سگ و گربه در محتوای مشابه آمده اند، تاثیر آن‌ها در تفسیر کلمه "jumped" نیز باید مشابه باشد (قاعدتا باید از کلمه ای مانند "پیراهن" متفاوت باشد چرا که پیراهن نمی‌پرد!). این ایده که کلماتی که در محتوای مشابه می‌آیند، از نظر معنایی نیز مشابه اند، نظریه ای بسیار قوی در معناشناسی لغوی[22] است.

معنا و معنا شناسی مفاهیم متفاوتی هستند. ایده ی معناشناسی لغوی توزیع شده[23] برای تعریف معنای کلمات توسط خواص توزیع شده (برای

[20] back-propagation
[21] likelihood
[22] lexical semantics
[23] distributionabl lexical semantics



شکل ۵.۳: تعبیه کلمه بر صفحه دو بعدی نگاشته شده است : کلمات از نظر معنایی مشابه، نزدیک به یکدیگر اند.[۶]

مثال محتوایی که در آن می آیند) آنهاست. کلماتی که در محتوای مشابه می آیند، نحوه نمایش مشابهی نیز دارند. در مدل برداری؛ مانند تعبیه کلمه ، تشابه بین لغات توسط معیارهایی مانند فاصله اقلیدسی یا فاصله کسینوسی بین زوایای بردارها اندازه گیری می‌شود.

اگر تعبیه کلمه‌های با ابعاد بزرگ را به صفحه دو بعدی بنگاریم، همانند شکل ۵.۳ مشاهده می‌شود که کلمات مشابه نزدیک به یکدیگر قرار می‌گیرند.

مدل‌های دیگری از مدل‌های زبانی نورونی وجود دارند که ما از ذکر آنها در این قسمت اجتناب کرده و به مقداری که توضیح داده شد، بسنده می‌کنیم.

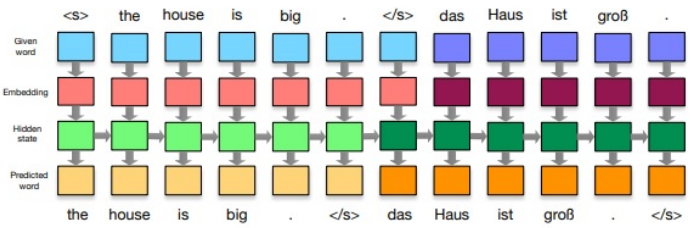
۴.۳ مدل ها ی ترجمه نورونی

۱.۴.۳ روش انکودر_ دیکدر

۲۴

نخستین تلاش در مدل ترجمه نورونی، تعمیمی از مدل زبان است. شبکه عصبی بازگشتی برای مدل سازی زبان بعنوان یک فرآیند پی در پی را به یاد

[24] Encoder-Decoder Approach



شکل ۶.۳: مدل انکودر_دیکدر جمله به جمله : برای گسترش مدل زبان، جمله ورودی the house is big را به معادل آلمانی آن، پیوند می دهیم.[۳]

آورید. با توجه به تمامی کلمات قبلی؛ این مدل، کلمه بعدی را پیش بینی می کند. وقتی به انتهای جمله می‌رسیم، ترجمه جمله را بصورت کلمه به کلمه پیش بینی می‌کنیم.

همانطور که در عکس ۳.۶ مشاهده می‌کنید، برای آموزش چنین مدلی، ما به سادگی جملات ورودی و خروجی را پیوند می‌دهیم و روشی را که برای آموزش یک مدل زبان استفاده کردیم، در اینجا به کار می‌گیریم. برای رمزگشایی، ورودی می‌دهیم و سپس پیش بینی را شروع می‌کنیم تا زمانی که آخرین توکن از جمله پیش بینی شود.

وقتی که عمل ترجمه به آخر جمله می‌رسد، حالت[۲۵] مخفی معنی آن را کد می‌کند. به برداری که مقادیر گرهای لایه مخفی را نگهداری می‌کند، جمله ورودی تعبیه شده[۲۶] گویند. این فاز، فاز انکدر مدل ترجمه است.در آینده از حالت مخفی برای ترجمه در فاز دیکدر استفاده می‌شود.[۱]

در طول فاز انکدر، لازم است تمامی اطلاعات جمله ورودی را بدانیم. از یاد بردن کلمه ابتدایی جمله در حالی که در حال پردازش انتهای جمله هستیم، پذیرفته نیست. در فاز دیکدر نه تنها برای پیش بینی هر کلمه به اطلاعات کافی نیاز داریم، بلکه باید بدانیم جمله تا کجا ترجمه شده و کجا هنوز ترجمه نشده است.[۱]

در عمل، مدل هایی پیشنهاد شدند که برای جملات کوتاه (۱۰ تا ۱۵ کلمه) نتایج معقولی داشته اند؛ اما، برای جملات بلندتر پاسخگو نیستند.

[25] state

[26] input sentence embedding

پس از بهبود این مدل‌ها، نتیجه به صورتی شد که در فاز دیکدر از حالت جمله تعبیه شده[۲۷] استفاده می‌شد که ورودی تمامی گره‌های لایه مخفی پیشین بود. این کار باعث می شد تا دیکدر از نظر ساختار با انکدر متفاوت باشد و سربار حالت مخفی را در طول کدگذاری کم کند؛ زیرا، نیازی به یادآوری کلمات ورودی نبود.

در قسمت بعد، درباره پشرفت چشم گیری از مدل صحبت خواهیم کرد : تراز کردن کلمات خروجی بر کلمات ورودی به صورت جداگانه.

۲.۴.۳ اضافه کردن مدل هم ترازی

در مدل ترجمه ماشینی نورونی از مدل انکدر_دیکدر توالی به توالی به همراه مکانیزم توجه استفاده می‌شود. این مدل؛ مشابه همان مدلی است که درباره اش صحبت شد، با این تفات که از یک مکانیزم هم ترازی در آن استفاده می‌شود که به آن 'توجه' گویند. ما در اینجا از کلمات هم ترازی و 'توجه' به جای هم استفاده می‌کنیم.

به دلیل این که مکانیزم 'توجه' مقداری مدل را پیچیده می‌کند، در ابتدا مروری می‌کنیم بر نحوه کار انکدر، دیکدر و سپس مکانیزم 'توجه' را مورد بررسی قرار می‌دهیم.

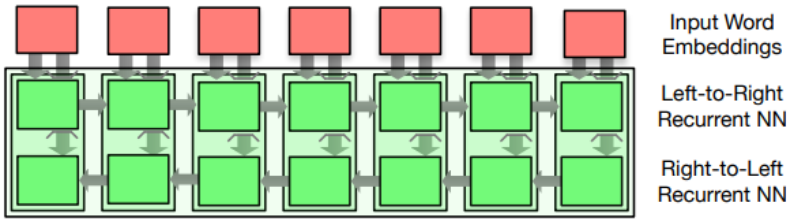
انکدر

وظیفه انکدر، ایجاد نمایشی برای جمله ورودی است. جمله ورودی دنباله ای از کلمات است. همان طور که در مدل‌های زبانی گفته شد، این کلمات توسط یک شبکه عصبی بازگشتی[۲۸] پردازش می‌شوند. این کار باعث ایجاد حالت های مخفی می شود که هر کلمه را با محتوای سمت چپ (کلمات قبل) آن رمزگذاری می‌کند. برای این که محتوای سمت راست نیز در نظر گرفته شود، یک شبکه عصبی بازگشتی نیز از راست به چپ؛ به طور دقیق تر از انتها به ابتدا، روی جمله اجرا می‌شود.

[27] sentence embedding state

[27] Alignment Model

[28] Recurrent Neural Network



شکل ۷.۳: مدل ترجمه ماشینی نورونی : انکدر ورودی. شامل دو عدد شبکه عصبی بازگشتی می‌شود که دو طرفه اجرا می‌شود. حالت ها ی انکدر ترکیبی از حالت های مخفی دو شبکه بازگشتی است.[۳]

شکل ۷.۳ این مدل را نشان می‌دهد. زمانی که دو شبکه عصبی بازگشتی داریم که از دو جهت اعمال می‌شوند، به آن شبکه عصبی بازگشتی دو طرفه[۲۹] گوئیم. انکدر شامل جستجوگر تعبیه شده برای هر کلمه ورودی x_j ، و نگاشتی که از طریق حالت های مخفی کار می کند، است.

$$\overleftarrow{h_j} = f(\overleftarrow{h_{j+1}}, \bar{E}x_j)$$

$$\overrightarrow{h_j} = f(\overrightarrow{h_{j-1}}, \bar{E}x_j)$$

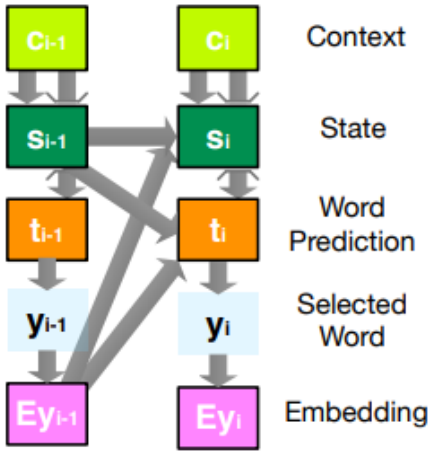
در این معادلات از تابع f برای هر سلول شبکه عصبی بازگشتی استفاده می‌کنیم. این تابع می‌تواند یک لایه شبکه عصبی پیش‌ـخور (مانند $f(x) = tanh(Ax + b)$) باشد یا پیچیده تر از آن.[۱]

در نظر داشته باشید که ما می‌توانیم با اضافه کردن گام پیش بینی کلمه در دنباله کلمات، این مدل را آموزش دهیم؛ اما، در اصل آن را در محتوای جمله آموزش می‌دهیم.

دیکدر

دیکدر نیز یک شبکه عصبی بازگشتی است. دیکدر، نمایش جمله ورودی، حالت های مخفی پیشین و پیش بینی کلمه خروجی را می‌گیرد و یک حالت رمزگشایی مخفی جدید و یک پیش بینی کلمه خروجی می‌سازد. (شکل ۸.۳)

[29] bidirectional RNN



شکل ۸.۳: مدل ترجمه ماشینی نورونی : دیکدر خروجی.[۳]

با شبکه عصبی بازگشتی که یک توالی از حالت های مخفی s_i را که توسط حالت مخفی پیشین s_{i-1} محاسبه شده، تعبیه کلمه خروجی قبلی E_{i-1} و محتوای ورودی c_i را نگه می دارد، شروع می‌کنیم.

$$s_i = f(s_{i-1}, Ey_{i-1}, c_i)$$

در اینجا نیز چندین انتخاب برای تابع f داریم : تبدیل خطی، GRU[۳۰] ، LSTM[۳۱] و ... . واضح است که تابع f در اینجا باید با f در انکدر همخوانی داشته باشد. یعنی اگر در انکدر از تبدیل خطی استفاده می‌کنیم، در دیکدر هم باید از تبدیل خطی استفاده شود. [۳]

حال از حالت مخفی، می‌خواهیم کلمه خروجی را پیش بینی کنیم.این پیش بینی به شکل یک توزیع احتمالی بر کل واژگان خروجی شکل می‌گیرد. اگر ۵۰۰۰۰ واژه داشته باشیم، بردار پیش بینی ۵۰۰۰۰ بعدی خواهد شد که هر درایه آن مربوط به احتمال پیش بینی شده برای هر کلمه در واژگان است.

بردار پیش بینی t_i ، حالت مخفی دیکدر s_{i-1} ، کلمه خروجی تعبیه شده

[30] gated recurrent units

[31] long short term memory

پیشین Ey_{i-1} و محتوای ورودی c_i را در بر می‌گیرد.

$$t_i = softmax(W(Us_{i-1} + VEy_{i-1} + Cc_i))$$

از softmax برای تبدیل بردار خام به یک توزیع احتمالی که مجموع تمام مقادیرش ۱ باشد، استفاده می‌شود. عموما، بیشترین مقدار بردار بیان کننده توکن خروجی y_i است.

در طول آموزش، خروجی درست y_i را می‌دانیم؛ در نتیجه، آموزش را با همین کلمه پیش می‌بریم. هدف آموزش این است که به کلمه خروجی درست، احتمال بیشتری بدهد. پس تابع هزینه آموزش باید منفی لگاریتم مقدار احتمال باشد.

$$cost = -\log t_i[y_i]$$

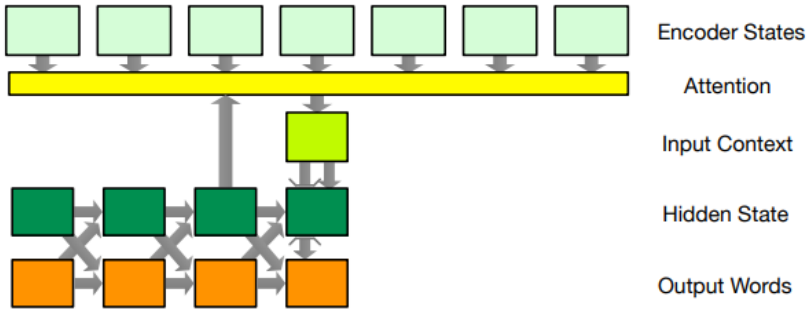
به طور ایده آل می‌خواهیم به کلمه درست، احتمال ۱ بدهیم که به معنی هزینه صفر است. ولی عموما هرچه احتمال کمتر باشد هزینه بالاتر است. توجه کنید که این تابع هزینه برای کلمات به صورت جداگانه است. یعنی هزینه کل جمله، مجموع هزینه‌های تک تک کلماتش است. [۳]

مکانیزم توجه

ما در حال حاضر، دو انتهای جمله داریم. دیکدر به ما نمایه کلمات $h_j = (\overleftarrow{h_j}, \overrightarrow{h_j})$ را می‌دهد و از ما انتظار دارد تا در هر گام i محتوای c_i را به او بدهیم. حال، مکانیزم 'توجه' را معرفی می‌کنیم که دو انتهای جمله را به هم متصل می‌کند.

نشان دادن مکانیزم 'توجه' توسط گراف ها دشوار است ولی شکل ۹.۳ شمایی از آن را به ما نشان می دهد. مکانیزم 'توجه' تمامی نمایه‌های کلمات ورودی $(\overleftarrow{h_j}, \overrightarrow{h_j})$ و حالت مخفی پیشین دیکدر s_{i-1} را می‌گیرد و یک حالت محتوایی c_i تولید می‌کند.

هدف این است که می‌خواهیم ارتباط بین حالت دیکدر (که به ما اطلاعاتی درباره ی محلی که در تولید جمله خروجی هستیم، می‌دهد) و هر کلمه ورودی را محاسبه کنیم. بر پایه این که این ارتباط چقدر قوی باشد؛ یا به عبارت دیگر، هر کلمه ورودی خاص چقدر به تولید کلمه خروجی بعدی مربوط است، می‌خواهیم تاثیر نمایه‌های کلمات را وزن دهی کنیم.[۲]



شکل ۹.۳: مدل ترجمه ماشینی نورونی : مدل 'توجه' . ارتباطات بین آخرین حال مخفی دیکدر و حالت ها ی انکدر محاسبه می‌شوند. از این ارتباطات برای محاسبه وزن‌های حالت ها ی انکدر استفاده می‌شود.[۳]

به صورت ریاضی، ابتدا این ارتباط را توسط یک لایه پیش‌ـخور محاسبه می‌کنیم (با استفاده از بردارهای وزن w^a و v^a و مقدار پایه b^a).

$$a(s_{i-1}, h_j) = w^{aT} s_{i-1} + u^{aT} h_j + b^a$$

خروجی این معادله یک مقدار عددی است که نشان می دهد کلمات ورودی چقدر برای تولید کلمه خروجی i مهم است.[۲]

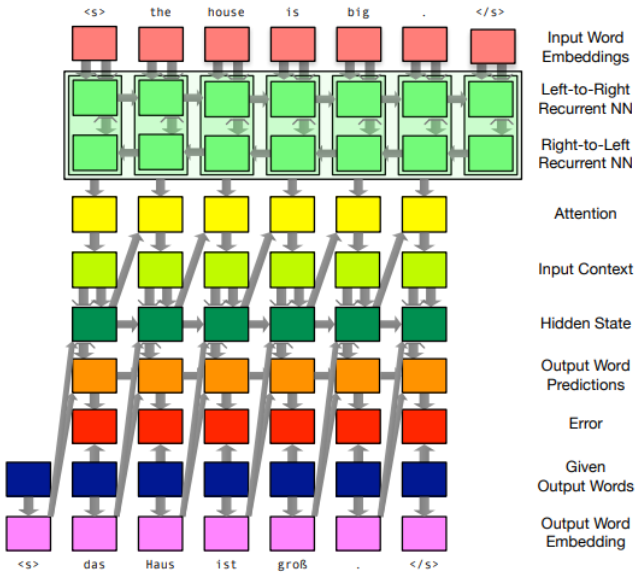
این مقدار 'توجه' را نرمال سازی می‌کنیم؛ در نتیجه، مقادیر 'توجه' بر روی تمامی کلمات ورودی j ، با استفاده از softmax ، حداکثر مقدار ۱ را اضافه می‌کند.

$$\alpha_{ij} = \frac{exp(a(s_{i-1}, h_j))}{\sum_k exp(a(s_{i-1}, h_j))}$$

حال از این مقدار نرمال شده برای وزن دهی به مشارکت کلمه ورودی h_j و بردار محتوای c_i استفاده می‌کنیم.

$$c_i = \sum_j \alpha_{ij} h_j$$

جمع بردارهای کلمات (وزن دار یا بی وزن) شاید در ابتدا بسیار عجیب و ساده به نظر برسد، اما در کاربرد روش‌های ترجمه ماشینی و یادگیری عمیق، بسیار شایع است.[۲]



شکل ۱۰.۳: یک گراف محاسباتی کاملا باز شده برای مثال آموزشی با ۷ توکن ورودی <s> the house is big </s> و ۶ توکن خروجی <s> grob das Haus is . تابع هزینه برای هر کلمه خروجی به طور جداگانه حساب شده و در طول جمله با هم جمع شده است.[۳]

۳.۴.۳ آموزش

حال با در دست داشتن مدل کامل، می‌توانیم نگاهی نزدیک تر به آموزش بیندازیم. یکی از چالش‌ها این است که برای داده‌های مختلف آموزشی، تعداد مراحل در انکدر با تعداد مراحل در دیکدر متفاوت است. زوج‌های جمله ای شامل جملات با طول متفاوت می‌شوند؛ در نتیجه، نمی‌توانیم برای همه داده‌های آموزشی یک گراف محاسباتی مشابه داشته باشیم. اما باید بتوانیم برای هر داده، به صورت پویا گراف محاسباتی بسازیم. به این تکنیک باز کردن[۳۲] شبکه عصبی بازگشتی گویند.

شکل ۱۰.۳ یک گراف محاسباتی کاملا باز شده برای محاسبه یک زوج

[31] Training
[32] unrolling

جمله کوتاه را نشان می‌دهد. در اینجا باید به چند نکته توجه داشته باشیم. خطایی که برای این زوج جمله محاسبه شده است، مجموع خطاهای هر کلمه است. وقتی به پیش بینی کلمه بعدی می‌رسیم، از کلمه صحیح برای محتوای حالت مخفی دیکدر و پیش بینی کلمه استفاده می‌کنیم. پس، هدف آموزش بر پایه ی احتمال داده شده به کلمه درست، با توجه به محتوای کامل است.

مدل‌های ترجمه ماشینی نورونی کاربردی، نیاز به GPU[۳۳] دارند؛ زیرا، که GPU ها برای حجم زیادی از محاسبات (مثلا ضرب تعداد زیادی از ماتریس ها) و موازی سازی‌ها در مدل‌های یادگیری عمیق بسیار مناسب اند. برای افزایش موازی سازی می‌خواهیم چندین زوج جمله (مثلا ۱۰۰ عدد) را یک‌جا پردازش کنیم. این کار باعث افزایش ابعاد تمامی تنسور[۳۴] های حالت می‌شود.

به عنوان مثال، ما هر کلمه ورودی در یک زوج جمله را با بردار h_j نشان می‌دهیم.چون ما از قبل یک توالی از کلمات ورودی داریم، این بردارها در یک ماتریس قرار می‌گیرند. سپس یک دسته از این زوج جملات را پردازش می‌کنیم. در نتیجه، ماتریس‌ها در یک تنسور سه بعدی قرار می‌گیرند.

به طور مشابه، حالت مخفی s_i دیکدر، به ازای هر کلمه یک بردار است. چون ما چندین جمله را پردازی می‌کنیم، این حالات مخفی در یک ماتریس قرار می‌گیرند.

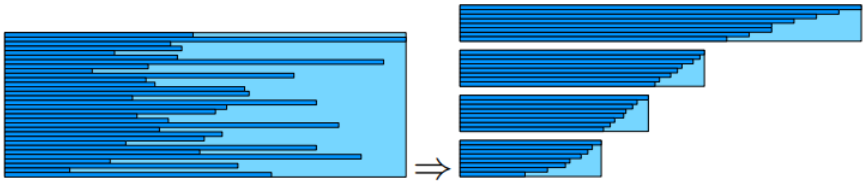
اولین مرحله ی محاسبه مکانیزم 'توجه' را به یاد آورید.

$$a(s_{i-1}, h_j) = w^a s_{i-1} + u^a h_j + b^a$$

می توانیم این محاسبه را در قالب یک ماتریس از حالت های s_{i-1} انکدر و یک تنسور سه بعدی از ورودی‌های کد شده s_j به GPU بدهیم. نتیجه آن یک ماتریس از مقادیر 'توجه' می شود (یک بعد برای زوج جمله و یک بعد برای کلمات ورودی). با توجه به استفاده بسیار زیاد از مقادیر W^a ، U^a و b^a ، همچنین موازی بودن ذاتی این محاسبات، GPU ها می توانند قدرت خود را نشان دهند. [۳]

[33] graphical process unit

[34] tensor



شکل ۱۱.۳: برای استفاده بهتر از موازی سازی در GPU ها ، یک دسته از دادهای آموزشی را همزمان باهم پردازش می کنیم. تبدیل یک دسته از دادهای آموزشی به دسته های کوچک تر با طول‌های یکسان، محاسبات را کمتر می‌کند.[۳]

حال ممکن است تصور کنید که تناقضی آشکار وجود دارد. ابتدا درباره ی این موضوع صحبت کردیم که هر داده آموزشی را باید در یک زمان انجام دهیم؛ زیرا که زوج جمله ها طول‌های متفاوت دارند و به گراف‌های محاسباتی با اندازه‌های مختلف نیاز است. سپس درباره ی پردازش موازی ۱۰۰ زوج جمله همزمان صحبت کردیم. این دو مسئله قطعا هدف‌های متناقضی هستند.

شکل ۱۱.۳ : وقتی دادها را دسته بندی می‌کنیم، باید بیشترین اندازه ورودی و خروجی را در یک دسته در نظر بگیریم و گراف محاسباتی را به مقدار بیشترین اندازه باز کنیم. برای جملات کوتاه تر، فاصله‌های باقیمانده را با غیر_ کلمات پر می‌کنیم و دنبال می‌کنیم که داده صحیح، کجا دارای پوشش[۳۵] است. به این معنی که باید اطمینان حاصل کنیم که هیچ 'توجه' ای بیشتر از طول جمله ورودی به کلمات داده نمی‌شود، و همچنین هیچ خطایی از کلمات خروجی بیشتر از طول جمله خروجی محاسبه نمی‌شود.

برای اجتناب از محاسبات اضافه فواصل جمله ها، می‌توانیم زوج جملات را به ترتیب طول شان مرتب کنیم و آنها را به دسته‌های کوچک تر با طول مساوی تقسیم کنیم.

آموزش؛ به طور خلاصه، شامل مراحل ذیل می شود :

- بر زدن پیکره دادهای آموزشی
- تفکیک پیکره به دسته‌های بزرگتر

[35] mask

- تفکیک هر دسته بزرگ به دستههای کوچکتر
- پردازش هر کدام از دستههای کوچکتر و جمع آوری گرادیانها
- اعمال تمام گرادیان ها به دستههای بزرگتر برای آپدیت پارامترها

عموما، آموزش یک مدل ترجمه ماشینی نورونی حدود ۵ تا ۱۵ اپوک[۳۶] (گذر هایی به تمام پیکره آموزشی) است. یکی از شاخصهای توقف، چک کردن پیشرفت مدل با استفاده از یک مجموعه اعتبار سنجی[۳۷] (که جزء داده آموزشی نیست) و توقف در صورت بهبود نیافتن خطا، است. آموزش طولانی تر لزوما نتیجه را بهبود نمیبخشد و حتی ممکن است باعث کاهش کارایی شود.

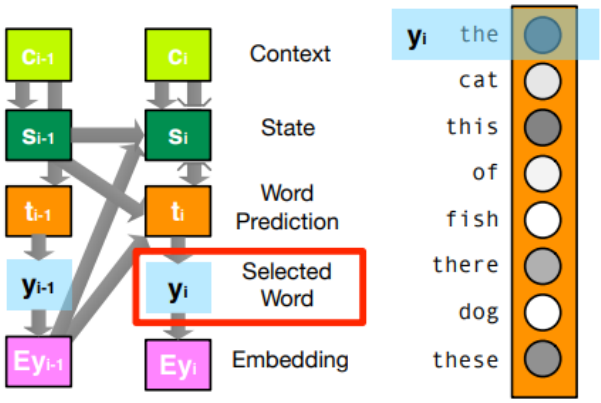
۴.۴.۳ جستجو ی شعاعی

عمل ترجمه با استفاده از مدلهای ترجمه ماشینی نورونی، یک گام در زمان انجام میشود. در هر گام، یک کلمه خروجی پیش بینی میشود. در مدل؛ ما، ابتدا یک توزیع احتمالی روی تمامی کلمات محاسبه میکنیم. سپس کلمه با بالاترین احتمال را انتخاب کرده و به گام بعدی میرویم. چون این مدل در شرایط کلمه خروجی قبلی بوده، از کلمه تعبیه شده آن میتوانیم برای شرایط محتوای مرحله بعد استفاده میکنیم. [۳] (شکل ۱۲.۳)

یک مثال واقعی از این که یک مدل ترجمه ماشینی نورونی، چطور یک جمله آلمانی را به انگلیسی ترجمه میکند، در شکل ۱۳.۳ نشان داده شده است. مدل تمایل دارد احتمال بیشتر را به انتخاب بالاتر بدهد اما ترجمه جمله نیز ابهام بیشتری به انتخاب کلمات میدهد؛ مانند think و believe یا different و various . همچنین ابهام درباره ی ساختار دستوری نیز وجود دارد؛ به طوری که جمله باید با but شروع شود یا با I .

در اینجا پیشنهاد میکنیم از "جستجو ی حریصانه یک ـ بهترین"[۳۸] استفاده شود. این کار ما را نسبت به مسئله garden-path آسیب پذیرتر میکند.

[36] epoch
[37] validation
[37] Beam Search
[38] 1-best greedy search



شکل ۱۲.۳: مرحله بتدایی رمزگشایی : مدل احتمال یک کلمه را پیش بینی می‌کند. محتمل ترین کلمه the را انتخاب می کنیم.[۳]

گاهی اوقات ما توالی از کلمات را دنبال می‌کنیم و اگر جایی اشتباه کرده باشیم، دیر متوجه آن می‌شویم. در این حالت، بهترین توالی آن است که در ابتدا کلمات محتمل کمتری دارد که توسط کلمات بعدی در متن خروجی استفاده می‌شود.

زمانی که اولین کلمه از کلمات خروجی پیش بینی شد، لیستی از تا n بهترین انتخاب ها را نگه می‌داریم و توسط احتمالشان، آن ها را امتیاز دهی می‌کنیم. سپس هر کدام از این کلمات را در محتوا ی کلمات بعد استفاده می‌کنیم. پیش بینی‌های متفاوتی به دست می‌آوریم. به واسطه این کار، امتیاز برای هر کلمه ترجمه شده و احتمالات کلمات پیش بینی شده را در هم ضرب می‌کنیم. زوج کلمه ای که بالاترین امتیاز را به دست آورد، برای لیست بعدی انتخاب می‌شود. (شکل ۱۴.۳)

این فرآیند ادامه می‌یابد. در هر گام زمانی، احتمالات ترجمه کلمات را نگه می‌داریم. وقتی ترجمه کامل شد، توکن "انتها ی جمله" تولید می‌شود. در این مرحله؛ لیست خالی شده و جستجو پایان می‌یابد.

جستجو، یک گراف فرضیات تولید می‌کند (شکل ۱۵.۳). از توکن "ابتدا ی جمله" $< s >$ شروع می‌کند و مسیرش در "انتها ی جمله" $< /s >$ خاتمه می‌یابد. ترجمه‌های حاصل می‌توانند از دنبال کردن نشانگرها بدست آیند.

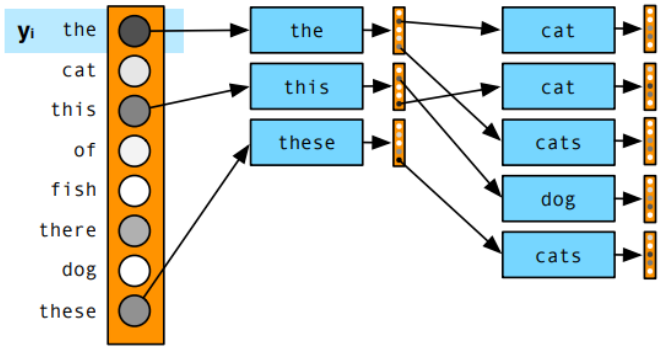
Input Sentence

ich glaube aber auch , er ist clever genug um seine Aussagen vage genug zu halten , so dass sie auf verschiedene Art und Weise interpretiert werden können .

Output Word Predictions

Best		Alternatives
but	(42.1%)	*however* (25.3%), *I* (20.4%), *yet* (1.9%), *and* (0.8%), *nor* (0.8%), ...
I	(80.4%)	*also* (6.0%), *,* (4.7%), *it* (1.2%), *in* (0.7%), *nor* (0.5%), *he* (0.4%), ...
also	(85.2%)	*think* (4.2%), *do* (3.1%), *believe* (2.9%), *,* (0.8%), *too* (0.5%), ...
believe	(68.4%)	*think* (28.6%), *feel* (1.6%), *do* (0.8%), ...
he	(90.4%)	*that* (6.7%), *it* (2.2%), *him* (0.2%), ...
is	(74.7%)	*'s* (24.4%), *has* (0.3%), *was* (0.1%), ...
clever	(99.1%)	*smart* (0.6%), ...
enough	(99.9%)	
to	(95.5%)	*about* (1.2%), *for* (1.1%), *in* (1.0%), *of* (0.3%), *around* (0.1%), ...
keep	(69.8%)	*maintain* (4.5%), *hold* (4.4%), *be* (4.2%), *have* (1.1%), *make* (1.0%), ...
his	(86.2%)	*its* (2.1%), *statements* (1.5%), *what* (1.0%), *out* (0.6%), *the* (0.6%), ...
statements	(91.9%)	*testimony* (1.5%), *messages* (0.7%), *comments* (0.6%), ...
vague	(96.2%)	*v@@* (1.2%), *in* (0.6%), *ambiguous* (0.3%), ...
enough	(98.9%)	*and* (0.2%), ...
so	(51.1%)	*,* (44.3%), *to* (1.2%), *in* (0.6%), *and* (0.5%), *just* (0.2%), *that* (0.2%), ...
they	(55.2%)	*that* (35.3%), *it* (2.5%), *can* (1.6%), *you* (0.8%), *we* (0.4%), *to* (0.3%), ...
can	(93.2%)	*may* (2.7%), *could* (1.6%), *are* (0.8%), *will* (0.6%), *might* (0.5%), ...
be	(98.4%)	*have* (0.3%), *interpret* (0.2%), *get* (0.2%), ...
interpreted	(99.1%)	*interpre@@* (0.1%), *constru@@* (0.1%), ...
in	(96.5%)	*on* (0.9%), *differently* (0.5%), *as* (0.3%), *to* (0.2%), *for* (0.2%), *by* (0.1%), ...
different	(41.5%)	*a* (25.2%), *various* (22.7%), *several* (3.6%), *ways* (2.4%), *some* (1.7%), ...
ways	(99.3%)	*way* (0.2%), *manner* (0.2%), ...
.	(99.2%)	*</s>* (0.2%), *,* (0.1%), ...
</s>	(100.0%)	

شکل ۱۳.۳: پیش بینی کلمات در یک مدل ترجمه ماشینی نورونی : عموما احتمال بیشتر به انتخاب بالاتر داده می‌شود اما کلماتی که از نظر معنایی مرتبط هستند، امتیاز بالاتری می گیرند. مانند believe با امتیاز ۶۸.۴ و کلمه think با امتیاز ۲۸.۶[۳]



شکل ۱۴.۳: جستجوی شعاعی در ترجمه ماشینی نورونی : پس از آن که لیستی از کلمات خروجی به دست آمد؛ برای هر کدام، پیش بینی ها ی جدیدی حاصل می‌شود.[۳]

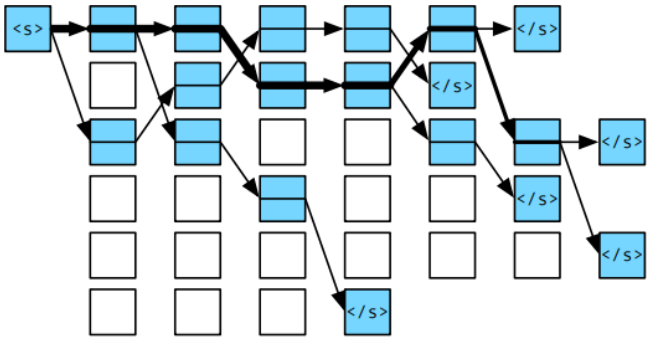
بالاترین امتیاز به بهترین ترجمه اشاره می‌کند.

زمانی که می‌خواهیم بهترین مسیر را انتخاب کنیم، هر مسیر را با ضرب احتمال‌های کلمات پیش بینی شده اش امتیاز دهی می‌کنیم. در عمل، نتیجه بهتر زمانی حاصل می‌شود که امتیاز را نسبت به طول ترجمه، نرمال سازی کنیم. این کار را پس از اتمام جستجو انجام می دهیم؛ زیرا، در حین جستجو، تمامی ترجمه‌های یک لیست طول یکسانی دارند. در نتیجه، نرمال سازی آنها تفاوتی ایجاد نخواهد کرد.

توجه کنید که در ترجمه آماری؛ اگر فرضیات، محتوای یکسان داشتند، می‌توانستیم آنها را ترکیب کنیم. این مسئله در شبکه‌های عصبی بازگشتی امکان پذیر نیست. در نتیجه، گراف جستجو در این مدل، تنوع کمتری نسبت به گراف جستجو در مدل آماری دارد.

۵.۳ جمع بندی

در این بخش، توضیحات کاملی دادیم درباره ی مدل اساسی ترجمه نورونی که امروزه بسیار رایج است. این مدل برای هر زوج زبانی، تقریبا خوب کار می‌کند.



شکل ۱۵.۳: گراف جستجو برای جستجوی شعاعی : در هر زمان، ۶ تا از بهترین ترجمه ها انتخاب می‌شوند. زمانی که به $< /s >$ برسیم، جمله خروجی کامل می‌شود. سپس لیست را کاهش می‌دهیم و زمانی که هر ۶ ترجمه کامل شدند، کار تمام است! [۳]

در طول این پروژه، درباره ترجمه ماشینی و به طور خاص؛ ترجمه ماشینی نورونی، شهودی مختصر پیدا کردیم. با روش انکدر_دیکدر آشنا شدیم که پایه و اساس تمامی روش‌های مختلف ترجمه نورونی هستند.همچنین دریافتیم ابزار هایی مانند google translate که کار بسیار دقیق و ظریفی را به لطف شبکه‌های عصبی، در کسری از ثانیه انجام می‌دهند، حجم زیادی دانش و ابتکار در وجود خود دارند.

Bibliography

[۱] Cho، Kyunghyun، et al. "On the properties of neural machine translation: Encoder-decoder approaches." arXiv preprint arXiv:۱۴۰۹.۱۲۵۹ .(۲۰۱۴)

[۲] Luong، Minh-Thang، Hieu Pham، and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:۱۵۰۸.۰۴۰۲۵ .(۲۰۱۵)

[۳] Koehn، Philipp. "Neural machine translation." arXiv preprint arXiv:۱۷۰۹.۰۷۸۰۹ .(۲۰۱۷)

[۴] Sennrich، Rico، et al. "Nematus: a toolkit for neural machine translation." arXiv preprint arXiv:۱۷۰۳.۰۴۳۵۷ .(۲۰۱۷)

[۵] Klein، Guillaume، et al. "Opennmt: Open-source toolkit for neural machine translation." arXiv preprint arXiv:۱۷۰۱.۰۲۸۱۰ .(۲۰۱۷)

[۶] Turian، Joseph، Lev Ratinov، and Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning." Proceedings of the ۴۸th annual meeting of the association for computational

linguistics. Association for Computational Linguistics، ۲۰۱۰.

[۷] MarianNMT : Fast Neural Machine Translation in C++ http://marian-nmt.github.io/

[۸] Forcada، Mikel L.، and Ramón P. Ñeco. "Recursive hetero-associative memories for translation." International Work-Conference on Artificial Neural Networks. Springer، Berlin، Heidelberg، ۱۹۹۷.

[۹] Castano، M. Asunción، Francisco Casacuberta، and Enrique Vidal. "Machine translation using neural networks and finite-state models." Theoretical and Methodological Issues in Machine Translation (TMI) (۱۹۹۷): ۱۶۰_۱۶۷.

واژه‌نامه

اپوک ..... epoch
احتمال ..... likelihood
اعتبار سنجی ..... validation
انتشار به عقب ..... back-propagation
آموزش ..... training
باروری ..... fertility
بازترکیبی ..... recombination
باز کردن ..... unrolling
پوشش ..... mask
نجزیه کننده تصادفی ..... stochastic parser
ترجمه ..... Translation
ترجمه ماشینی ..... Machine Translation
ترجمه ماشینی آماری ..... Statistical Machine Translation
ترجمه ماشینی قاعده مند ..... Rule-Based Machine Translation
ترجمه ماشینی مبتنی بر انتقال ..... Transfer-Based Machine Translation
ترجمه ماشینی مبتنی بر عبارت ..... Phrase-Based Machine Translation
ترجمه ماشینی مبتنی بر کلمه ..... Word-Based Machine Translation
ترجمه ماشینی مبتنی بر متن ..... Corpus-Based Machine Translation
ترجمه ماشینی مبتنی بر مثال ..... Example-Based Machine Translation
ترجمه ماشینی مبتنی بر نحو ..... Syntax-Based Machine Translation
ترجمه ماشینی نورونی ..... Neural Machine Translation
تطبیق الگو ..... Pattern Matching

تطبیق ماژول ..... Module Matching
تنسور ..... tensor
تئوری بیز ..... Bayes Theory
جستجوی شعاعی ..... Beam search
جیتجوی حریصانه یک بهترین ..... 1-best greedy search
جمله ورودی تعبیه شده ..... input embedding sentence
حالت ..... state
حالت جمله تعبیه شده ..... embedding sentence state
خوشه بندی ..... clustering
روش انکودر_دیکدر ..... Encoder-Decoder Approach
شبکه عصبی ..... Neural Network
شبکه عصبی بازگشتی ..... Recurrent Neural Network
شبکه عصبی پیچشی ..... Convolutional Neural Network
لایه مخفی ..... hidden layer
عبارت اسمی ..... Noun Phrase
عبارت فعلی ..... Verbal Phrase
متن ..... Corpus
مدل توالی_به_توالی ..... Sequence-to-sequence Model
مدل چند لایه ..... Multi layers Model
مدل خطی ..... Linear Model
معناشناسی لغوی ..... lexical semantics
معناشناسی لغوی توزیع شده ..... distribiutional lexical semantics
میان زبانی ..... Interlingua
نظریه اطلاعات ..... Information Theory
هموار سازی ..... Smoothing
یادگیری ماشین ..... Machine Learning
GPU ..... Graphical Process Unit

Abstract

The life of human beings depends on communicating with each other. Since long time ago، due to communication، man has been able to conquer his difficulties. Modern man wants to be in touch with his surroundings and has a powerful tool to overcome the issue of having multiple languages all over the world and his short lifetime to learn all of them. Translation، an essential technique for humanity.

The objective of translation is to bring better understanding and create a deeper relationship among people around the globe during the history، with any language، race or culture. Translators have always had a great deal of service. Nowadays، with the advancements of computer science، computers are supposed to be future translators.

In this project، we will review a variety of machine translation methods. First، we explain the traditional methods that have long been used، then we describe the neural machine translation method which is a revolution in machine translation.

Faculty of Science
School of mathematics، statistics and computer science

A Survey On Neural Machine Translation

By

Shaghayegh Yousefpour

Supervisor

Dr. Bagher BabaAli

Project for recieving bachelor degree
Computer Science

July 2018