
بسم الله الرحمن الرحيم

Learning, Natural Language Processing

چکیده به زبان فارسی:

یادگیری ژرف رویکرد جدیدی در حوزه‌ی یادگیری ماشین است. یادگیری ژرف دیدگاه نوینی برای ترجمه‌ی ماشینی فراهم کرده است. این دیدگاه به ترجمه‌ی ماشینی عصبی معروف است. آموزش یکپارچه‌ی این سیستم‌ها برای ترجمه‌ی ماشینی انگلیسی به فارسی، نیاز به پیکره‌ی متنی موازی بزرگ دارد. متاسفانه در زبان فارسی چنین پیکره‌ی متنی بزرگی در دسترس نیست. به عنوان یک راه‌حل جایگزین، پیکره‌ی متنی بزرگ فارسی، برای بازنمایی برداری کلمات جمع‌آوری شد. برای بازنمایی برداری کلمات از روش Word2Vec استفاده شده است. مجموعه آزمونی برای ارزیابی بازنمایی برداری کلمات با استفاده از معیار شباهت کسینوسی تعریف شد. با توجه به نتایج به دست آمده از آزمایش‌ها، پارامترهایی که در بازنمایی برداری کلمات به روش Word2Vec تاثیر داشتند عبارتند از: حجم مجموعه داده‌ها و بُعد بردار ویژگی. هر چه بُعد بردار بالاتر و حجم مجموعه داده‌ها بیشتر باشد، بازنمایی بهتری خواهیم داشت. برای پیاده‌سازی سیستم ترجمه‌ی ماشینی عصبی از مدل دنباله به دنباله استفاده شد و ورودی سیستم ترجمه را با بازنمایی‌های حاصل از روش Word2Vec، مقداردهی اولیه کردیم. آموزش مدل پیشنهادی برای ترجمه، 45 ساعت طول کشید و پس از آموزش مدل به سرگشتگی 29 رسید.

چکیده به زبان انگلیسی:

Deep learning is a new area of machine learning research. Deep learning has proposed a new approach for machine translation which is called neural machine translation. End-to-end training of this system on the task of English–to-Persian translation needs a large parallel corpus. There is a few parallel corpora for Persian language which is either small in size or unavailable for research purpose. As an alternative solution, large Persian

corpus for vector representation of words has been collected. Word2Vec has been used for vector representation of words. To measure quality of the Persian word vectors, we defined a new test set that contains four types of sematic questions, and six types of syntactic questions. In order to answer these questions, we used cosine similarity measure. Our results show that the accuracy of model depends on dimension of the word vectors and amount of data. In this study, we used the recently proposed sequence to sequence framework for neural machine translation. The vector representation of words from Word2vec were used in order to initializing the neural machine translation system. The proposed model have been taken 2 days to train and it achieved a perplexity of 29.

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

چکیده

یادگیری ژرف رویکرد جدیدی در حوزه‌ی یادگیری ماشین است. یادگیری ژرف دیدگاه نوینی برای ترجمه‌ی ماشینی فراهم کرده است. این دیدگاه به ترجمه‌ی ماشینی عصبی معروف است. آموزش یکپارچه‌ی این سیستم‌ها برای ترجمه‌ی ماشینی انگلیسی به فارسی، نیاز به پیکره‌ی متنی موازی بزرگ دارد. متاسفانه در زبان فارسی چنین پیکره‌ی متنی بزرگی در دسترس نیست. به عنوان یک راه‌حل جایگزین، پیکره‌ی متنی بزرگ فارسی، برای بازنمایی برداری کلمات جمع‌آوری شد. برای بازنمایی برداری کلمات از روش Word2Vec استفاده شده است. مجموعه آزمونی برای ارزیابی بازنمایی برداری کلمات با استفاده از معیار شباهت کسینوسی تعریف شد. با توجه به نتایج به دست آمده از آزمایش‌ها، پارامترهایی که در بازنمایی برداری کلمات به روش Word2Vec تاثیر داشتند عبارتند از: حجم مجموعه داده‌ها و بُعد بردار ویژگی. هر چه بُعد بردار بالاتر و حجم مجموعه داده‌ها بیشتر باشد، بازنمایی بهتری خواهیم داشت. برای پیاده‌سازی سیستم ترجمه‌ی ماشینی عصبی از مدل دنباله به دنباله استفاده شد و ورودی سیستم ترجمه را با بازنمایی‌های حاصل از روش Word2Vec، مقداردهی اولیه کردیم. آموزش مدل پیشنهادی برای ترجمه، ۴۵ ساعت طول کشید و پس از آموزش مدل به سرگشتگی ۲۹ رسید.

واژگان کلیدی: Word2Vec،بازنمایی برداری کلمات، ترجمه‌ی ماشینی عصبی، یادگیری ژرف، پردازش زبان طبیعی

فصل ۱

کلیات پژوهش

۱.۱ مقدمه

ترجمه‌ی ماشینی[۱]، فرآیند آموزش به ماشین‌ها است تا بتوانند ترجمه خودکار بین زبان‌های مختلف را یاد بگیرند و از این‌رو حوزه‌ی مهمی در امر پژوهش تلقی می‌شود. ایده‌های فلسفی ابتدایی زبان‌های فراگیر[۲] در قرن هفدهم تا اولین پیشنهادهای کاربردی در دهه‌ی ۵۰ میلادی نشان می‌دهد که ترجمه‌ی ماشینی تاریخچه‌ای طولانی دارد [۲۹]. یکی از تاثیرگذارترین این پیشنهادها توسط ویور[۳] ارائه شد [۷۵]. پیشنهاد ویور سرآغاز پژوهش‌های ترجمه‌ی ماشینی در ایالت متحده آمریکا شد. ویور در پیشنهادش قصد داشت تا با ترکیب دانش‌ها در زمینه‌های آمار[۴]، رمزنگاری[۵] و نظریه اطلاعات[۶] و همچنین قوانین فراگیر زبانشناسی و منطق، از کامپیوترها برای ترجمه (به خصوص با در نظر گرفتن مسئله ابهام زبان)

[۱] Machine Translation (MT)

[۲] universal languages

[۳] Weaver

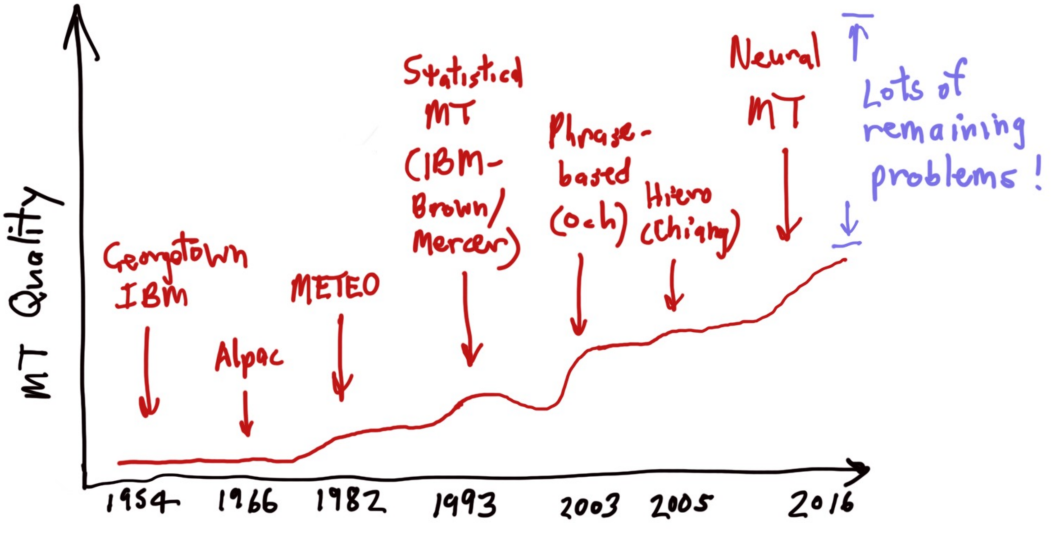
[۴] statistics

[۵] cryptography

[۶] information theory

استفاده کند [۲۸]. پس از آن، ترجمه‌ی ماشینی در دوره‌های زیادی پیشرفت چشمگیری داشت. در شکل ۱.۱ روند پیشرفت ترجمه‌ی ماشینی آورده شده است. در این پژوهش از رویکرد ترجمه‌ی ماشینی عصبی[۷] استفاده می‌شود که ورودی سیستم ترجمه، بردارهایی هستند که از روش Word2Vec به دست آمده‌اند. ترجمه‌ی ماشینی عصبی یک رویکرد جدید برای ترجمه‌ی ماشینی است که در طول دو سال اخیر توسعه داده شده است. نتایج به دست آمده از کار پژوهشگران روی ترجمه‌ی ماشینی عصبی باعث پیشرفت قابل‌توجهی در کیفیت ترجمه‌ی ماشینی شده است (شکل ۱.۱).

در این فصل ابتدا به تاریخچه‌ی ترجمه‌ی ماشینی پرداخته می‌شود. سپس به تعریف مساله و اهمیت و ضرورت آن و همچنین اهداف و پرسش‌های پژوهش خواهیم پرداخت. در انتهای فصل نیز ساختار فصل‌های باقی‌مانده را شرح می‌دهیم.



شکل ۱.۱: روند پیشرفت ترجمه‌ی ماشینی [۴۰]

[۷] Neural Machine Translation (NMT)

۲.۱ تاریخچه‌ی ترجمه‌ی ماشینی

ترجمه‌ی ماشینی در سال‌های ۱۹۶۰-۱۹۵۰، بیشتر بر اساس جایگزینی کلمه به کلمه با استفاده از فرهنگ لغت دوزبانه بود. دوره افول ترجمه‌ی ماشینی دقیقا بعد از گزارش ALPAC[۸] در سال ۱۹۶۶ شروع شد. این گزارش بیان می‌کرد: «هیچ چشم‌انداز قابل پیش‌بینی و نزدیکی برای استفاده از ترجمه‌ی ماشینی وجود ندارد» [۲۹] که همین امر مانع پژوهش‌های ترجمه‌ی ماشینی بیش از یک دهه شد. با تجدید حیات ترجمه‌ی ماشینی در آغاز دهه‌ی ۸۰ میلادی از کشورهای اروپایی، ژاپن و کمی بعد در ایالت متحده آمریکا، ترجمه‌ی ماشینی آماری[۹] توسط محققان IBM در سال ۱۹۹۳ شروع شد [۸]. رویکردهای مبتنی بر پیکره‌ی متنی[۱۰] پیشنهاد شده، به محتوای زبانی کمتری نیاز داشت و برای آموزش سیستم‌های ترجمه‌ی ماشینی فقط به یک مجموعه داده‌های موازی از زوج جمله‌هایی که ترجمه‌ی یکدیگرند، نیاز بود. در این رویکرد به جای ساخت فرهنگ لغت دوزبانه که کاری زمان‌بر بود، براون[۱۱] و همکارانش مدلی ارائه دادند که این فرهنگ لغت یا مدل‌های ترجمه را به صورت احتمالاتی از پیکره‌های متنی موازی[۱۲] یاد می‌گرفتند. سپس از این سیستم برای ترجمه‌ی جملات جدید، استفاده می‌شد (شکل ۲.۱). برای انجام این کار، براون و همکارانش ۵ الگوریتم، معروف به مدل‌های ۱ تا ۵ IBM ارائه دادند. این الگوریتم‌ها برای یادگیری هم‌ترازی[۱۳] کلمات با ایجاد یک نگاشت بین کلمات زبان مبدا[۱۴] و زبان مقصد[۱۵] در یک پیکره‌ی متنی موازی، استفاده می‌شد (شکل ۳.۱). ایده این روش

[۸] Automatic Language Processing Advisory Committee

[۹] Statistical Machine Translation (SMT)

[۱۰] corpus

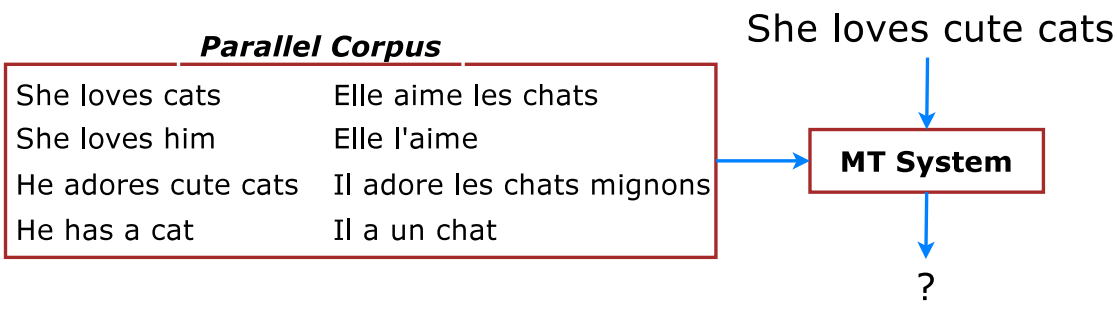
[۱۱] Brown

[۱۲] parallel corpora

[۱۳] alignment

[۱۴] source language

[۱۵] target language

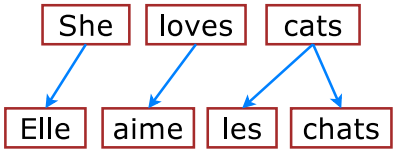


شکل ۲.۱: رویکرد مبتنی بر پیکره‌ی متنی موازی برای ترجمه‌ی ماشینی

ساده است: هر چه دو کلمه مثل "love" و "aime" در زوج جمله‌های متفاوتی با هم بیایند، احتمال بیشتری وجود دارد که این دو با هم هم‌تراز بوده و معنای یکسانی داشته باشند.

تعریف پیکره‌ی متنی

پیکره متنی مجموعه‌ای بزرگ و بدون ساختار از متون تولید شده توسط انسان است. از پیکره برای آموزش یا ارزیابی مدل‌های پردازش زبان طبیعی استفاده می‌شود. پیکره می‌تواند یک زبانه یا چند زبانه باشد. در صورتی‌که در پیکره چند زبانه ارتباطی بین جملات زبان‌های مختلف تعریف شده باشد، به آن پیکره موازی می‌گویند که کاربرد آن بیشتر در ترجمه‌ی ماشینی است.



شکل ۳.۱: هم‌ترازی مبتنی بر کلمه

هنگامی‌که یک مدل ترجمه، مثل یک فرهنگ لغت دوزبانه احتمالاتی، یاد گرفته شد، مدل IBM ۱ که ساده‌ترین الگوریتم از بین ۵ الگوریتم معرفی شده است؛ یک جمله مبدا جدید را به این صورت ترجمه می‌کند: