

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

چکیده به زبان فارسی:

یادگیری ژرف رویکرد جدیدی در حوزه یادگیری ماشین است. یادگیری ژرف دیدگاه نوینی برای ترجمه ماشینی فراهم کرده است. این دیدگاه به ترجمه ماشینی عصبی معروف است. آموزش یکپارچه این سیستم‌ها برای ترجمه ماشینی انگلیسی به فارسی، نیاز به پیکره‌ی متنی موازی بزرگ دارد. متأسفانه 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 ترجمه 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 semantic 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

ت	فهرست شکل‌ها
ج	فهرست جدول‌ها
۱	۱ کلیات پژوهش
۱	۱.۱ مقدمه
۳	۲.۱ تاریخچه‌ی ترجمه‌ی ماشینی
۹	۳.۱ تعریف مساله
۱۰	۴.۱ اهمیت و ضرورت پژوهش
۱۱	۵.۱ اهداف و پرسش‌های پژوهش
۱۲	۶.۱ ساختار پایان‌نامه
۱۳	۲ مبانی نظری ترجمه‌ی ماشینی عصبی
۱۳	۱.۲ مدل زبانی
۱۴	۱.۱.۲ مدل N-gram
۱۵	۲.۱.۲ مدل‌های زبانی مبتنی بر شبکه‌های عصبی
۱۷	۳.۱.۲ ارزیابی مدل‌های زبانی
۱۸	۲.۲ شبکه‌ی عصبی بازگشتی
۲۱	۱.۲.۲ مدل‌های زبانی مبتنی بر شبکه‌های عصبی بازگشتی

۲۳	یادگیری بر پایه پس انتشار خطا	۲.۲.۲
۲۵	الگوریتم پس انتشار خطا برای شبکه‌ی عصبی بازگشتی	۳.۲.۲
۳۰	LSTM	۳.۲
۳۸	آموزش LSTM	۱.۳.۲
۳۸	الگوریتم پس انتشار خطا برای LSTM	۱.۱.۳.۲
۴۰	ترجمه‌ی ماشینی عصبی	۴.۲
۴۲	معماری رایج شبکه‌های عصبی بازگشتی برای ترجمه‌ی ماشینی	۱.۴.۲
۵۱	بازنمایی برداری کلمات	۳
۵۱	یادگیری ژرف	۱.۳
۵۳	یادگیری بازنمایی	۱.۱.۳
۵۳	بازنمایی برداری کلمات	۲.۳
۵۵	بازنمایی برداری پیوسته کلمات	۱.۲.۳
۵۹	بازنمایی برداری کلمات به روش Word2Vec	۳.۳
۶۰	معماری CBOW	۱.۳.۳
۶۶	معماری skip-gram	۲.۳.۳
۶۹	روش پیشنهادی و ارزیابی و نتایج آن	۴
۶۹	روش پیشنهادی	۱.۴
۷۲	مجموعه داده‌ها	۲.۴
۷۲	پیکره‌ی همشهری دو	۱.۲.۴
۷۲	پیکره‌ی موازی انگلیسی-فارسی میزان	۲.۲.۴
۷۴	پیکره‌ی موازی انگلیسی-فارسی تهران	۳.۲.۴
۷۵	پیش پردازش مجموعه داده‌ها	۴.۲.۴
۷۶	پارامترهای مهم یادگیری در روش Word2Vec	۳.۴

۷۷	ارزیابی بازنمایی برداری توزیع شده کلمات فارسی	۴.۴
۸۱	نتایج و تحلیل بازنمایی برداری توزیع شده کلمات فارسی	۵.۴
۸۵	پیاده‌سازی مدل دنباله به دنباله برای ترجمه‌ی ماشینی عصبی	۶.۴
۸۷	نتایج سیستم ترجمه‌ی ماشینی عصبی	۱.۶.۴
۸۹		نتیجه‌گیری و کارهای آینده	۵
۹۰	پیشنهادها و کارهای آینده	۱.۵
۹۱		واژه‌نامه فارسی به انگلیسی	
۹۶		واژه‌نامه انگلیسی به فارسی	
۱۰۱		مرجع‌ها	

فهرست شکل‌ها

۲	روند پیشرفت ترجمه‌ی ماشینی [۴۰]	۱.۱
۴	رویکرد مبتنی بر پیکره‌ی متنی موازی برای ترجمه‌ی ماشینی	۲.۱
۴	هم‌ترازی مبتنی بر کلمه	۳.۱
۵	یک مدل ساده از فرآیند ترجمه توسط الگوریتم IBM ۱	۴.۱
۷	ترجمه‌ی ماشینی مبتنی بر عبارت	۵.۱
۹	معماری کدکننده-کدگشا برای ترجمه‌ی ماشینی	۶.۱
۱۸	شبکه‌ی عصبی بازگشتی ساده [۵۹]	۱.۲
	نمایی از یک شبکه‌ی عصبی بازگشتی که در گام‌های زمانی بسط پیدا کرده است	۲.۲
۱۹	[۵۹]	
۲۰	نمایی از یک شبکه‌ی عصبی بازگشتی ساده و وزن‌های بین لایه‌ها	۳.۲
۲۲	نمونه‌ای از مدل زبانی بازگشتی	۴.۲
۲۵	تاثیر نرخ یادگیری بر روند آموزش به کمک الگوریتم گرادینان کاهش [۳۱]	۵.۲
	واحدهای تکرار شونده در شبکه‌های عصبی بازگشتی استاندارد که فقط دارای یک	۶.۲
۳۲	لایه هستند [۵۹]	
۳۲	واحدهای تکرار شونده در LSTM ها که دارای ۴ لایه هستند [۵۹]	۷.۲
	سلول حافظه را می‌توان به صورت یک تسمه نقاله تصور کرد که از اول تا آخر	۸.۲
۳۳	دنباله با تعاملات خطی جزئی در حرکت است [۵۹]	

۹.۲	یک دریچه که از واحد سیگموئیدی به همراه یک عملگر ضرب نقطه به نقطه تشکیل شده است [۵۹].	۳۳
۱۰.۲	دریچه‌ی فراموشی که تصمیم می‌گیرد که مقدار قبلی سلول فراموش یا حفظ شود [۵۹].	۳۴
۱۱.۲	دریچه‌ی ورودی که تصمیم می‌گیرد مقدار از ورودی به سلول حافظه وارد شود یا خیر [۵۹].	۳۵
۱۲.۲	به‌روزرسانی سلول حافظه در LSTM [۵۹]	۳۶
۱۳.۲	دریچه‌ی خروجی که تصمیم می‌گیرد چه اطلاعاتی را به خروجی ببرد [۵۹].	۳۶
۱۴.۲	واحد بازگشتی گیتی [۵۹]	۴۲
۱۵.۲	مدل دنباله-به-دنباله (seq2seq) برای ترجمه‌ی ماشینی عصبی	۴۳
۱۶.۲	مثالی از این که چگونه یک مدل ترجمه‌ی ماشینی عصبی آموزش دیده، جمله ورودی را با الگوریتم کدگذاری حریصانه ترجمه می‌کند.	۴۹
۱.۳	بازنمایی برداری کلمات به روش one-hot [۵۴]	۵۴
۲.۳	معماری 4-gram مدل زبانی مبتنی بر شبکه‌ی عصبی پیشرو	۵۷
۳.۳	نمایی کلی از روش Word2Vec	۵۹
۴.۳	معماری ساده‌ی CBOW با زمینه‌ی تک کلمه‌ای [۶۴]	۶۱
۵.۳	معماری CBOW با زمینه‌ی چند کلمه‌ای [۶۴]	۶۶
۶.۳	معماری skip-gram [۶۴]	۶۷
۱.۴	معماری پیشنهادی برای ترجمه‌ی ماشینی عصبی با استفاده از بازنمایی برداری توزیع شده Word2Vec	۷۱
۲.۴	بخشی از پیکره‌ی همشهری دو	۷۳
۳.۴	بخشی از پیکره‌ی موازی انگلیسی-فارسی میزان	۷۳
۴.۴	بخشی از پیکره‌ی موازی انگلیسی-فارسی TEP	۷۴

۵.۴	مقایسه دقت دو مدل skip-gram و CBOW برای قسمت فارسی مجموعه داده‌ی
۸۳	TEP
۶.۴	مقایسه دقت دو مدل skip-gram و CBOW برای قسمت فارسی مجموعه داده‌ی
۸۴	میزان

فهرست جدول‌ها

۷۴	جزئیات پیکره‌ی موازی انگلیسی-فارسی میزان	۱.۴
۷۵	جزئیات پیکره‌ی موازی انگلیسی-فارسی تهران	۲.۴
۳.۴	مثال‌هایی از سوال‌های معنایی و نحوی از مجموعه آزمون فارسی برای ارزیابی	
۷۹	مدل‌های CBOW و skip-gram	
۸۲	نتایج حاصل از ارزیابی قسمت فارسی مجموعه داده‌ی TEP در مدل skip-gram	۴.۴
۸۲	نتایج حاصل از ارزیابی قسمت فارسی مجموعه داده‌ی TEP در مدل CBOW	۵.۴
۸۳	نتایج حاصل از ارزیابی قسمت فارسی مجموعه داده‌ی میزان در مدل skip-gram	۶.۴
۸۴	نتایج حاصل از ارزیابی قسمت فارسی مجموعه داده‌ی میزان در مدل CBOW	۷.۴
۸۵	نتایج حاصل از الحاق سه مجموعه داده‌ی همشهری دو، میزان و TEP	۸.۴
۸۷	نمونه‌هایی از ترجمه‌های نامناسب تولید شده توسط مدل پیشنهادی	۹.۴
۸۸	نمونه‌هایی از ترجمه‌های مناسب تولید شده توسط مدل پیشنهادی	۱۰.۴

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. متأسفانه 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 پیشنهادها^۳ 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

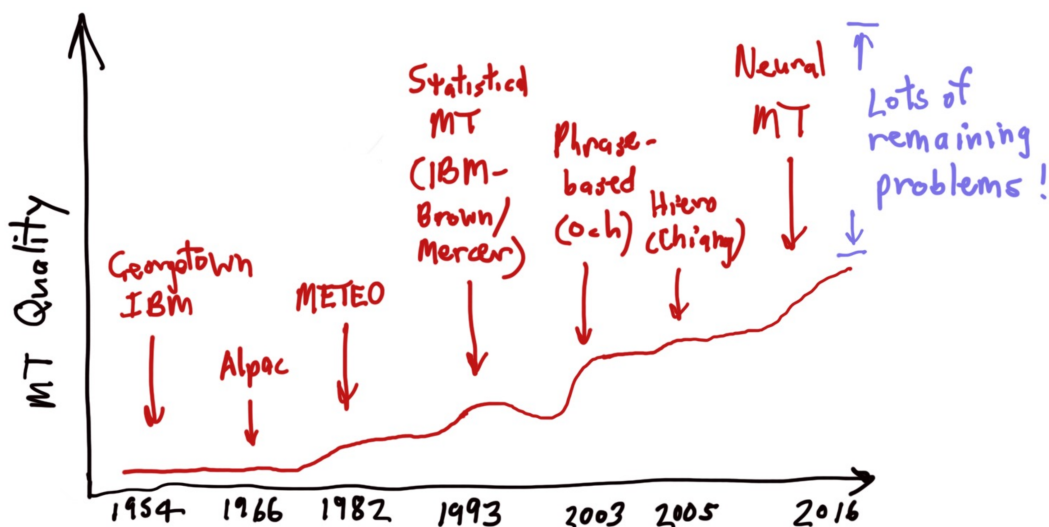
^۴ statistics

^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 قابل توجهی 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]

^۷Neural 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 دقیقاً 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 مبدأ^{۱۴} و زبان مقصد^{۱۵} در یک پیکره‌ی متنی موازی، استفاده می‌شد (شکل ۳.۱). ایده این روش

^۸Automatic Language Processing Advisory Committee

^۹Statistical Machine Translation (SMT)

^{۱۰} corpus

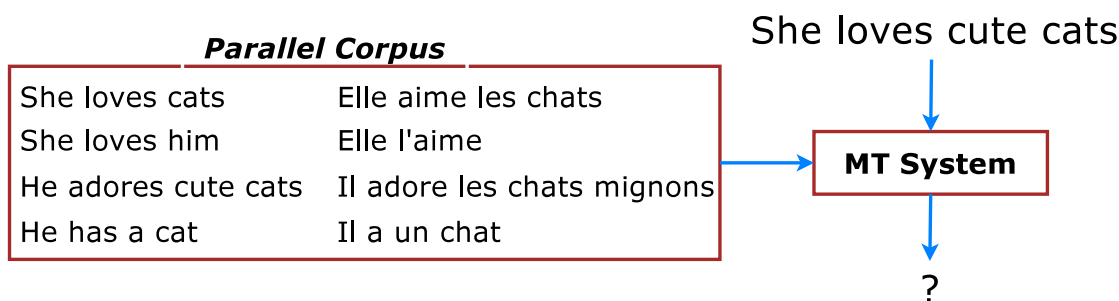
^{۱۱} Brown

^{۱۲} parallel corpora

^{۱۳} alignment

^{۱۴} source language

^{۱۵} target language

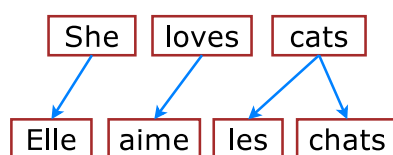


شکل ۲.۱: رویکرد مبتنی بر پیکره‌ی متنی موازی برای ترجمه‌ی ماشینی

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 زبان یا چند زبان باشد. در صورتی که در پیکره چند زبان 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

هنگامی که 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: