دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

گزارش پروژه‌ی درس پردازش زبان‌های طبیعی

موضوع:

«غنی‌سازی متن به کمک شعر، غزل، نقل‌قول و حکمت»

اعضای تیم:

مهدی سعیدی، فرانک کریمی، محمدرضا کمالی، آرمان مظلوم‌زاده

استاد درس:

دکتر احسان‌الدین عسگری

اسفند ۱۴۰۱

چکیده

موضوع غنی‌کردن متن یکی از مباحث جذاب و چالش‌برانگیز در بین موضوعات مطرح در پردازش زبان طبیعی است که در کاربردهایی مانند زبان‌شناسی مورد استفاده قرار می‌گیرد. اما این کار همواره با چالش‌هایی مانند استخراج معنا از عبارات به صورت غیروابسته به زبان و جمع‌آوری داده‌های مناسب روبرو بوده است. برای استخراج معنا از عبارات، مدل‌های مختلفی وجود دارند که ارزیابی آن‌ها در زبان‌های مختلف امری دشوار است. بدین منظور در پژوهش حاضر، در ابتدا داده‌هایی با ساختارهای مختلف (مانند شعر، نقل‌قول و حکمت) به زبان‌های پارسی، انگلیسی و عربی جمع‌آوری شده‌اند و سپس با استفاده از مدل‌های پیشرو در ساخت امبدینگ، کار غنی‌کردن متن ورودی به کمک این مجموعه داده‌ها و امبدینگ مربوط به آن‌ها انجام شده است. در آخر برای محقق‌کردن هدف این پژوهش، خروجی مدل‌های مختلف برای ورودی‌های مشخص بررسی و به صورت تحت نظارت برچسب‌گذاری شده‌اند تا در مرحله‌ی بعدی بتوان مدل‌های مورد استفاده را ارزیابی کرد. در نتیجه‌ی این ارزیابی، مدل LaBSE با دقتی معادل ۸۰ درصد در معیار NDCG، عملکرد قابل‌قبولی داشته است و می‌توان گفت برای داده‌های با ساختارهای مختلف و به صورت چندزبانه قابل‌استفاده است.

مقدمه

در این پژوهش مجموعه‌ای از پرس‌وجوهای ورودی در اختیار داریم که جملاتی به زبان فارسی هستند و قصد داریم خروجی‌های مدل‌های مختلف مورد نظر را در قبال این ورودی‌ها بررسی کنیم و به آن‌ها نمره دهیم. در آخر هم این نمره‌دهی را به کمک روش‌های علمی ارزیابی کنیم و با انجام ارزیابی مدل‌ها را با یکدیگر مقایسه کنیم.

در خصوص کارهای مشابه با وجود اینکه کار کاملا مشابه پیدا نشد، ولی پژوهشی هست که نزدیک به پژوهش گفته‌شده به حساب می‌آید. در این پژوهش ([۳])، یک رویکرد برای زبان‌شناسان ارائه شده است که بتوانند داده‌های موجود را با داده‌های دیگر موجود در وب طی سه مرحله غنی کنند. این مراحل نهایتا به ساخت نمایه‌های زبانی منجر می‌شود که علاوه بر داده‌هایی که در مراحل میانی تولید می‌شود، جستجو بر روی آن‌ها هم می‌تواند انجام شود. اما چیزی که پژوهش حاضر را نسبت به این پژوهش خاص می‌کند، استفاده از چندین مدل مختلف است که برای تولید امبدینگ جملات به کار گرفته شده‌اند همچنین، در مرحله‌ی ارزیابی این پژوهش یک ارزیابی تحت نظارت و به صورت علمی انجام می‌شود که عملکرد مدل‌های مختلف را با یکدیگر مقایسه می‌کند و نهایتا بهترین مدل را معرفی می‌کند.

روش‌ها

در این بخش داده‌ای که جمع‌آوری شده است به همراه روش جمع‌آوری آن شرح داده شده است. سپس به معرفی مدل‌هایی که برای تولید امبدینگ‌ها از آن‌ها استفاده شده است پرداخته می‌شود. نهایتا آزمایش‌ها و ارزیابی‌هایی که انجام شده‌اند به همراه تقسیم کار آورده می‌شود.

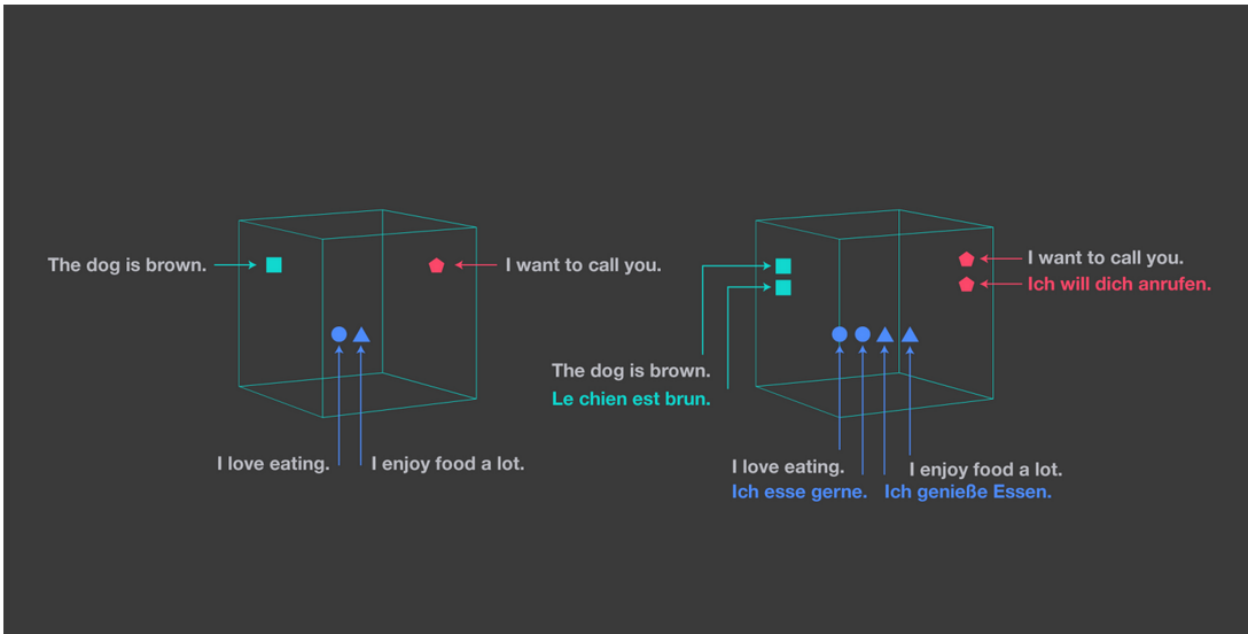
داده‌ها و نحوه جمع‌آوری آن

داده‌هایی که در این پژوهش برای غنی‌سازی مورد استفاده قرار گرفته‌اند شامل ۳۲۲۹ غزل از دیوان شمس مولوی، ۴۹۵ غزل حافظ و ۳۵۰ تک‌بیتی از دیگر اشعار زبان فارسی است. همچنین ۴۸۰ حکمت از نهج‌البلاغه به زبان عربی و ۲۰۴۰ نقل‌قول به زبان انگلیسی نیز در مجموعه‌ی داده قرار گرفتند تا پژوهش به صورت چندزبانه انجام شود. نحوه‌ی جمع آوری غزل‌های دیوان شمس و حافظ از فایل‌های با فرمت doc بوده است که از [۵] دریافت شدند و به شکل مناسب درآورده شدند. نقل‌قول‌های انگلیسی هم از مجموعه‌داده‌ی

[۶] بدست آمده است که نقل‌قول‌های منتخب کاربران سایت گودریدز از کتاب‌های انگلیسی‌زبان است. استخراج تک‌بیتی‌ها یک توسط یکی از اعضای تیم به صورت دستی انجام شده است. این مجموعه تک‌بیتی‌ها یکی از خروجی‌های این پژوهش به حساب می‌آید. برای استخراج این مجموعه از منابع [۱۰] –[۷] استفاده شده است. حکمت‌های نهج‌البلاغه هم از مجموعه‌داده‌ی معرفی‌شده در درس بدست آمده است.

مدل‌های مورد استفاده

در این پژوهش، پس از بررسی مدل‌های مختلف و با در نظر گرفتن اینکه نیاز به مدلی داریم که امکان امبدینگ چندزبانه را داشته باشد، مدل‌های LaBSE [۱]، LASER [۴] و FastText (Tf-idf weighted) [۲] انتخاب شدند تا مورد استفاده قرار بگیرند. این مدل‌ها در این پژوهش به ما کمک می‌کنند تا متون مدنظر خود را به یک فضای برداری مشترک ببریم که بدین واسطه شباهت بین متون به کمک روش‌هایی مانند شباهت کسینوسی محاسبه شود. این مدل‌های کلمات هم‌معنای زبان‌های مختلف را به بردارهای نزدیک به هم نگاشت می‌کنند. تصویر ۱ این موضوع را به خوبی نشان می‌دهد.



تصویر ۱–نمایشی از فضای امبدینگ چندزبانه

آزمایش‌ها و ارزیابی‌ها به همراه تقسیم کار

کاری که در این پژوهش انجام شده است را می‌توان به بخش‌های «جستجو و جمع‌آوری داده»، «جستجو و انتخاب مدل‌ها»، «انتخاب مجموعه داده ورودی»، «خروجی گرفتن از مدل‌ها»، «ایجاد محیط تحت وب»، «استقرار محیط تحت وب و مدل روی سرور»، «برچسب‌گذاری جفت‌های ورودی خروجی»، «بررسی صحت برچسب‌گذاری» و «ارزیابی عملکرد مدل‌ها» تقسیم کرد. در ابتدای انجام پژوهش، سه کار اول به شکل همزمان بین اعضا تقسیم شدند و پس از آن سه کار بعدی شروع به انجام شد. نهایتا در روزهای انتهایی کار برچسب‌گذاری انجام شد و تمامی زوج‌های ورودی خروجی ممکن توسط افراد تیم برچسب خوردند. مقادیر این برچسب‌ها یکی از مقادیر صحیح بین ۲– تا ۲+ بود. برچسب‌های منفی برای مشخص کردن تضاد مفاهیم بین زوج‌های ورودی و خروجی استفاده می‌شود و برچسب‌های مثبت برای نشان دادن هم‌راستایی دو متن ورودی و خروجی کاربرد دارد. مقدار صحیح برچسب میزان همبستگی معنایی

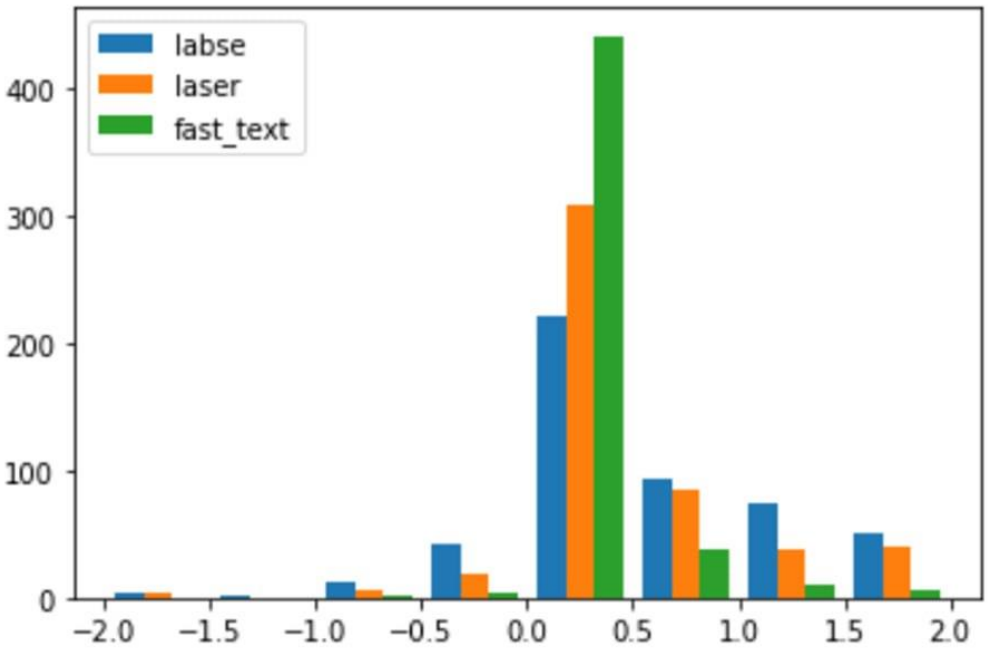
ورودی و خروجی را نشان می‌دهد. به این ترتیب عدد ۲ به زوجی که بیشترین تشابه معنایی را دارند تخصیص داده می‌شود و عدد ۲- برای زوج‌هایی که دو مفهوم کاملاً متضاد را نشان می‌دهند، به کار می‌رود و عدد صفر برای زوج‌های ورودی و خروجی با مفهوم نامرتبط است. برای سنجش صحت برچسب‌گذاری، از روش Fleiss' kappa استفاده شد که در نتیجه‌ی آن به عدد ۰.۲۹ رسیدیم که یعنی توافق رای‌دهندگان «منصفانه» بوده است و بنابراین این برچسب‌گذاری علمی است. سپس ارزیابی نهایی بر روی عملکرد مدل‌ها در مواجهه با پرس‌وجوهای تعریف‌شده مطابق برخی از معیارهای موجود در [۱۱] انجام شد که مهم‌ترین آن‌ها معیار NDCG (Normalized Discounted Cumulative Gain) است.

جدول ۱- برچسب‌های انتخاب شده برای ارزیابی ورودی و خروجی

مفهوم برچسب	برچسب
تضاد کامل معنایی بین ورودی و خروجی	۲-
تضاد نسبی معنایی بین ورودی و خروجی	۱-
عدم همبستگی معنایی بین ورودی و خروجی	۰
هم‌راستایی نسبی مفهوم بین ورودی و خروجی	۱+
هم‌راستایی کامل مفهوم بین ورودی و خروجی	۲+

نتایج

پس از اینکه به صورت تحت نظارت برچسب‌ها زده شدند، یک توزیع به شکلی که در تصویر ۲ قابل‌مشاهده است به دست آمد.



تصویر ۲- توزیع برچسب‌ها در خروجی مدل‌های مختلف

همچنین نتایج ارزیابی‌های نهایی انجام‌شده در این پژوهش در جداول ۱ تا ۳ آمده است. توجه داشته باشید که با توجه به اینکه در این پژوهش از مجموعه حکمت‌های نهج‌البلاغه به‌ازای هر مدل و هر ورودی تنها یک خروجی داشتیم، معیار NDCG همواره برابر یک می‌شود و بنابراین در جدول ۱ این مجموعه داده آورده نشده است.

جدول ۲ – ارزیابی مدل‌ها بر اساس NDCG

	LaBSE	LASER	FastText (Tf-idf weighted)
تک‌بیتی	0.8	0.72	0.38
نقل‌قول	0.8	0.78	0.16
غزلیات شمس	0.8	0.6	0.2
غزلیات حافظ	0.78	0.5	0.26

جدول ۳ – ارزیابی مدل‌ها بر اساس CG@1

	LaBSE	LASER	FastText (Tf-idf weighted)
تک‌بیتی	0.43	0.38	0.13
نقل‌قول	0.58	0.55	0.06
غزلیات شمس	0.52	0.3	0.07
غزلیات حافظ	0.54	0.2	0.1
نهج‌البلاغه	0.72	0.39	0.03

جدول۴ – ارزیابی مدل‌ها بر اساس MRR

	LaBSE	LASER	FastText (Tf-idf weighted)
تک‌بیتی	0.6	0.21	0.05
نقل‌قول	0.52	0.29	0.02
غزلیات شمس	0.54	0.17	0.03
غزلیات حافظ	0.46	0.09	0.06
نهج‌البلاغه	0.4	0.24	0.001

لازم به ذکر است که در انتهای این پژوهش یک وب‌سایت[1] نیز ایجاد شده است که با استفاده از Milvus[2] امکان جستجوی یک عبارت را فراهم می‌کند که در نتیجه‌ی آن مواردی که برای غنی‌کردن آن مناسب هستند، در قالب دسته‌های جداگانه (غزل حافظ، تک‌بیتی، حکمت و...) به کاربر ارائه می‌شود.

نتیجه‌گیری

مطابق نتایج بدست‌آمده از ارزیابی انجام‌شده، مدلی که در تمامی معیارها عملکرد بهتری داشته LaBSE است که توانسته در تمامی مجموعه‌های داده‌ای و با زبان‌های مختلف عملکرد قابل‌قبولی از خود به جای گذارد. در ادامه‌ی این پژوهش، تعداد مجموعه‌های داده‌ای افزایش خواهد یافت و مجموعه‌های داده‌ای مانند ضرب‌المثل و کنایه نیز به داده‌های غنی‌کننده اضافه می‌شود. همچنین برای بدست‌آوردن امبدینگ مدل‌هایی نظیر GPT (با حذف لایه آخر) را نیز اضافه خواهیم کرد و عملکرد آن‌ها روی مجموعه‌ی داده‌ای جدید مجدداً ارزیابی می‌شود. همچنین با توجه به ضعف FastText در به کارگیری Tf-idf، باید بررسی‌های بیشتری در صحت به‌کارگیری آن انجام شود و نهایتا در صورت لزوم با به‌کارگیری روش‌های دیگر عملکرد آن بهبود داده شود.

[1] https://nlp.armanexplorer.com
[2] https://github.com/milvus-io/milvus

مراجع

[1] Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W., 2020. Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852.

[2] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, pp.135-146.

[3] Xia, F. and Lewis, W., 2009, March. Applying NLP technologies to the collection and enrichment of language data on the web to aid linguistic research. In Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH–SHELT&R 2009) (pp. 51-59).

[4] https://github.com/facebookresearch/LASER

[5] https://bigdata-ir.com/

[6] https://huggingface.co/datasets/Abirate/english_quotes

[۷] نحوی، محمد، «تک‌بیت‌های ناب از سخنوران پارسی‌گوی کم‌نام(جلد اول)» ، تهران: انتشارات رنگینه/ ۱۳۹۹

[۸] نحوی، محمد، «تک‌بیت‌های ناب از سخنوران پارسی‌گوی کم‌نام(جلد دوم)» ، تهران: انتشارات رنگینه/ ۱۳۹۹

[9] https://roozaneh.net/

[10] https://setare.com

[11] https://amitness.com/2020/08/information-retrieval-evaluation