به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس پردازش زبان طبیعی

استاد ممتازی

تمرین سوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: ابهام‌زدایی معنایی کلمات

پیش از ارائه نتایج نکات کلی در مورد پیاده‌سازی و مجموعه‌داده را بیان می‌کنم:

- مجموعه آموزش اولیه را به دو قسمت اعتبارسنجی (۲۰٪) و آموزش (۸۰٪) شکاندم تا از مجموعه اعتبارسنجی برای تعیین ابرپارامترهای مدل دسته‌بند استفاده کنم.
- پیاده‌سازی BERT استفاده‌شده تنها می‌تواند برای کلمات یک پنجره کوچک تعبیه ارائه دهد و مابقی کلمات را دور می‌ریزد؛ لذا متن کامل به مدل داده نمی‌شود و تنها چند کلمه قبل و چند کلمه به مدل داده می‌شود و بردار کلمه هدف به دسته‌بند داده می‌شود.
- برخی از داده‌ها دارای خطا است؛ یعنی ممکن است مفهوم نامرتبطی برای یک کلمه مبهم ارائه شده باشد. این متون چه در آموزش و چه در ارزیابی کنار گذاشته می‌شود. در جدول زیر مشخص شده است که برای شکل‌های مختلف یک کلمه مختلف چه مفاهیم در نظر گرفته شده است:

کلمه	اشکال مختلف کلمه	مفاهیم کلمه
Hard	hard, harder, hardest	HARD1, HARD2, HARD3
Interest	interest, interested, interesting, interests	Interest1, interest2, interest3, interest4, interest5, interest5
Line	line, lines	cord, division, formation, phone, product, text
Serve	serve, served, serves	SERVE10, SERVE12

- برای هر کلمه مبهم یک SVM جداگانه آموزش داده شده است که کارش تشخیص مفهوم آن کلمه مبهم است.
- برای هر SVM از کرنل‌های مختلف خطی، چندجمله‌ای و RBF استفاده شده است و بهترین دسته‌بند بر اساس صحت روی مجموعه اعتبارسنجی انتخاب شده است.

- برای کلمه Hard تنها مفهوم HARD1 در مجموعه آموزش آمده است؛ لذا دسته‌بند SVM نمی‌توان برای آن آموزش داد. در این شرایط به ازای تمام داده‌ها همان مفهوم HARD1 پیش‌بینی می‌شود.
- معیارهای دقت به ازای هر کلمه مبهم ارائه شده است و در نهایت با میانگین وزن‌دار دقت‌های نهایی نیز ارائه شده است.

آمار مربوط به هر کلمه در هر مجموعه‌داده در جدول زیر آورده شده است:

کلمه	تعداد داده آموزش	تعداد داده اعتبارسنجی	تعداد داده آزمون
Hard	۱۳۳۴	۳۲۷	۳۰۰
Interest	۱۳۴۳	۳۳۷	۲۸۵
Line	۱۱۸۵	۲۷۰	۲۳۶
Serve	۱۱۱۲	۳۰۴	۲۷۶
کلی	۴۹۷۴	۱۲۳۸	۱۰۹۷

بهترین تنظیم برای هر کلمه در جدول زیر آورده شده است:

کلمه	بهترین تنظیم
Hard	-
Interest	کرنل خطی
Line	کرنل RBF
Serve	کرنل خطی

صحت هر دسته‌بند و صحت کلی در جدول زیر آورده شده است:

کلمه	مجموعه آموزش	مجموعه اعتبارسنجی	مجموعه آزمون
Hard	۱۰۰/۰۰٪	۱۰۰/۰۰٪	۱۰۰/۰۰٪
Interest	۱۰۰/۰۰٪	۹۱/۶۹٪	۸۹/۹۲٪
Line	۹۷/۳۸٪	۹۲/۵۹٪	۹۴/۴۹٪

Serve	۱۰۰/۰۰٪	۱۰۰/۰۰٪	۱۰۰/۰۰٪
کلی	۹۹/۳۸٪	۹۶/۱۲٪	۹۶/۱۷٪

دقت F1 هر دسته‌بند و دقت F1 کلی در جدول زیر آورده شده است:

کلمه	مجموعه آموزش	مجموعه اعتبارسنجی	مجموعه آزمون
Hard	۱۰۰/۰۰٪	۱۰۰/۰۰٪	۱۰۰/۰۰٪
Interest	۱۰۰/۰۰٪	۶۲/۹۶٪	۷۴/۶۱٪
Line	۹۷/۲۰٪	۹۲/۷۴٪	۹۴/۰۹٪
Serve	۱۰۰/۰۰٪	۱۰۰/۰۰٪	۱۰۰/۰۰٪
کلی	۹۹/۳۳٪	۸۸/۳۳٪	۹۲/۱۳٪

بخش دوم: ایجاد تجزیه‌کننده روابط وابستگی

برای انجام این تمرین و پوشش بخش امتیازی آن تنظیمات مختلفی را بررسی کرده‌ام:

- برای تعبیه کلمات دو مدل Word2Vec و GloVe بررسی شده است.
- برای لایه از بازگشتی هم از BiLSTM و هم از BiGRU ارزیابی شده است.
- دو شبکه مورد استفاده بوده است: یک شبکه ساده متشکل از تنها یک لایه بازگشتی و یک لایه خروجی و شبکه پیشنهادی پیچیده‌تری شامل دو لایه بازگشتی پشته‌شده، دو لایه متراکم، یک لایه Dropout و نهایتا یک لایه خروجی

نکات مهم در مورد پیاده‌سازی من عبارت است از:

- از ۲۰ گام آموزش به همراه یک کالبک EarlyStopping استفاده شده است تا مدل فرصت داشته باشد به اندازه کافی آموزش ببیند در عین حال اگر شرایط پیش‌برازش به وجود آمده باشد، آموزش خاتمه بیابد.
- برای مدل‌های تعبیه کلمه از بردارهای ۳۰۰ بعدی استفاده شده است.
- دنباله‌های کوتاه‌تر حاشیه‌گذاری شده‌اند تا تمام دنباله‌ها هم اندازه باشد.
- به دلیل وجود توکن‌های حاشیه‌ای معیارهای ارزیابی تغییر یافته است تا تاثیر این توکن‌ها را حذف کند. باتوجه به اینکه حاشیه‌گذاری تعداد توکن زیادی را اضافه می‌کند، تاثیر نامطلوب آن زیاد است. به عنوان مثال اگر صحت پیش‌فرض ۹۵٪ بوده است، صحت پس از حذف کلاس حاشیه‌ای و محاسبه صحت واقعی به حدود ۸۰٪ کاهش پیدا می‌کند.
- باتوجه به حجیم بودن داده‌ها از مفهوم مولد مجموعه‌داده استفاده کرده‌ام تا رم کمتری استفاده شود.

در جدول زیر نتایج برای ۸ مدل پیشنهادی آورده شده است:

مدل تعبیه	سلول بازگشتی	ساختار شبکه	صحت آموزش	صحت اعتبارسنجی	صحت آزمون
Word2Vec	LSTM	ساده	۸۰/۵۳٪	۷۷/۱۲٪	۷۷/۳۶٪
Word2Vec	LSTM	پیچیده	۸۵/۸۸٪	۸۱/۲۹٪	۸۱/۴۴٪
Word2Vec	GRU	ساده	۷۷/۲۰٪	۷۴/۷۸٪	۷۵/۵۶٪
Word2Vec	GRU	پیچیده	۷۸/۲۳٪	۷۶/۱۲٪	۷۶/۶۱٪
GloVe	LSTM	ساده	۷۸/۴۶٪	۷۵/۶۲٪	۷۵/۹۸٪
GloVe	LSTM	پیچیده	۷۵/۶۰٪	۷۳/۴۳٪	۷۳/۷۴٪
GloVe	GRU	ساده	۷۱/۵۷٪	۶۹/۴۱٪	۷۰/۲۶٪
GloVe	GRU	پیچیده	۷۶/۲۲٪	۷۴/۱۳٪	۷۴/۸۵٪

اولین سطر این جدول مربوط به حالت پیش‌فرض است که در صورت سوال خواسته شده است و سایر تنظیمات برای بهبود آن ارائه شده است. همانطور که مشخص است با تنظیم پیش‌فرض می‌توان به صحت آزمون ۷۷/۳۶٪ رسید. چنانچه از شبکه پیچیده (دو لایه بازگشتی، دو لایه متراکم و یک لایه Dropout) استفاده شود و مدل تعبیه و نوع سلول بازگشتی به ترتیب همان Word2Vec و LSTM باشد می‌توان به صحت بهتری معادل ۸۱/۴۴٪ دست پیدا کرد که بهترین صحت است.

از نتایج بر می‌آید که به طور کلی تعبیه Word2Vec بهتر از تعبیه GloVe، شبکه متشکل از سلول‌های LSTM بهتر از شبکه متشکل از سلول‌های GRU و شبکه پیچیده ارائه‌شده بهتر از شبکه ساده ارائه‌شده است.

در ادامه برای تحلیل و ارزیابی نتایج بر روی بهترین مدل (سطر دوم) گزارش خواهد شد:

الف) این معیارها در جدول زیر آورده شده است. چون مدل از نوع دسته‌بند چندکلاسه است، برای محاسبه Precision و Recall باید نتایج کلاس‌های مختلف تجمیع شود. اگر متناسب با اندازه کلاس‌ها وزن‌دهی صورت بگیرد دقت این دو معیار برابر با

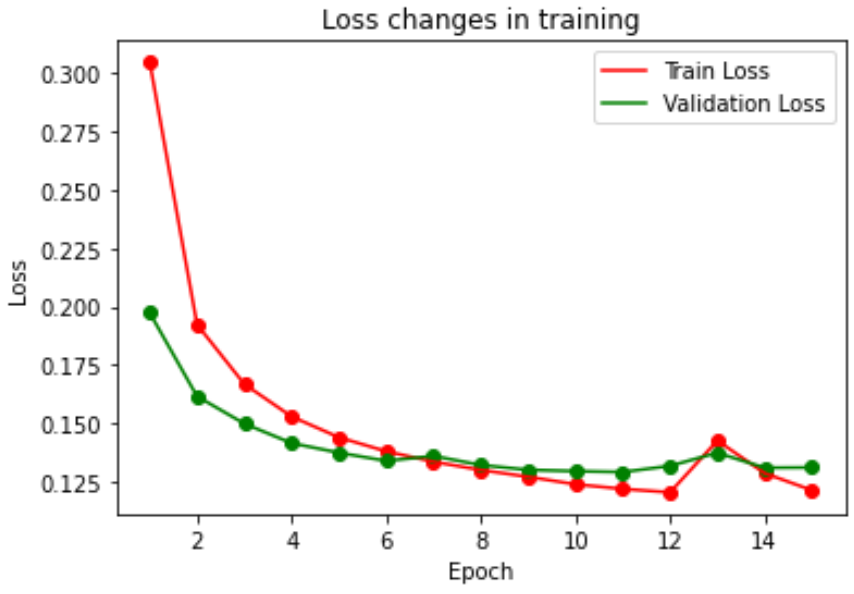
Accuracy می‌شود؛ پس برای اینکه اعداد دیگری گزارش شود، میانگین غیر وزن‌دار کلاس‌های مختلف محاسبه و گزارش شده است:

F1	Recall	Precision	Accuracy
۲۵/۳۱٪	۱۸/۴۸٪	۴۰/۱۱٪	۸۰/۲۱٪

مقدار Accuracy مطابق با انتظار بالاست ولی Precision و Recall مقادیر خیلی کمی را دارد؛ با توجه به اینکه میانگین غیروزن‌دار کلاس‌ها محاسبه شده است، تمام کلاس‌ها تاثیر یکسانی در این معیارهای دقت داشته‌اند. به وضوح برخی از کلاس‌ها مانند Root، 1L و 1R بسیار پرتکرار است و مدل روی آن دقت خوبی دارد ولی برخی از کلاس‌ها مانند 50L به ندرت دیده می‌شود و شاید مدل ترجیح بدهد چنین خروجی را هیچ وقت تولید نکند. وجود این کلاس‌های کم تکرار این دو معیار را کاهش داده است.

علت اینکه Precision از Recall بیشتر است هم به دلیل آن است که در Precision برخی از کلاس‌ها اصلا تولید نشده است و از میانگین حذف شده است ولی در Recall تمام کلاس‌ها وجود دارند و تاثیر کلاس‌های اصلی بیشتر کم شده است.

ب)



ج) برای سه جمله مذکور به ترتیب داریم: