به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس پردازش متن و زبان طبیعی

سحر رجبی
شماره دانشجویی
۸۱۰۱۹۹۱۶۵

اردیبهشت ماه 1400

LSTM-4

گام‌های پیاده‌سازی:

بعد از انجام پیش‌پردازش‌های مربوطه (کوچک‌کردن حروف، تعیین جملات، حذف علائم نگارشی، جایگزین کردن کلماتی که در glove نیستند با <UNK> و ...) باید کلمات را به نحوی کدگزاری کنیم که توسط مدل قابل درک باشد. در نتیجه به هر کلمه یک index اختصاص می‌دهیم و دو dictionary که می‌تواند کلمات را به ایندکس، و ایندکس را به کلمات مپ کند تشکیل می‌دهیم؛ سپس تمامی کلمات را با ایندکس متناظر عوض می‌کنیم.

مرحله‌ی بعدی تعیین ورودی و خروجی سیستم است. ما می‌خواهیم که یک sequene to sequence مدل

داشته‌باشیم. به این صورت که هر sequenceای از واژگان، که به مدل داده شود؛ مدل کلمه‌ی بعد آن را پیش‌بینی کند. همچنین یکی از مزایای مدل‌های مبتنی بر lstm، این است که می‌توانیم سایز رشته‌ها را متفاوت در نظر بگیریم. با توجه به این توضیحات، برای هر جمله، کلمه‌ی اول تا یکی مانده به آخر به عنوان ورودی مدل و کلمه‌ی دوم تا آخر را به عنوان خروجی درنظر میگیریم. (در واقع زمانی که به صورت رشته به آنها نگاه کنیم، خروجی کلمه‌ی بعد از هر کلمه است).

یکی از مشکلات برای آموزش شبکه طول متفاوت sequenceهاست. زمانی که در یک batch می‌خواهیم تعداد جمله را به مدل بدهیم، طول آن‌ها باید ثابت باشد. برای حل این مشکل دو راه حل وجود دارد: ۱- جملات با طول یکسان را در یک batch قرار دهیم. ۲- از padding استفاده کنیم.

من در ابتدا از روش اول استفاده کردم که در کمال تعجب باعث کاهش خطا در داده‌های test نشد. اما مراحل این روش به این صورت انجام شده: در ابتدا داده‌ای با طول یکسان را در یک گروه قراردادیم که البته، تعداد اعضای گروه‌های با طول جملات کوتاه، بسیار بیشتر از طول جملات بلند بود؛ و از آنجایی که هر mini batch یک بار دیده خواهد شد؛ اگر ما از این دسته‌ها به این صورت استفاده کنیم، در حالی که learning rate ثابتی داریم؛ تاثیر یک جمله با طول زیاد بسیار بیشتر از یک جمله با طول کم خواهد شد. برای رفع این مشکل، من حداکثر اندازه‌ی batch را برابر ۱۲۸ قرار دادم. یعنی دسته‌های با تعداد اعضای بیشتر خود به دسته‌های کوچکتر تقسیم خواهند شد. یک راه دیگر برای رفع این مشکل، آن است که learning rate ما بر اساس طول دسته انتخاب شود و به این صورت میزان تاثیرات را کنترل کنیم. همانطور که گفته‌شد با وجود این راه حل‌ها، ما نتیجه‌ی مناسبی از این روش نگرفتیم.

در ادامه با استفاده از padding این مشکل حل شد. در این روش، باید کاری کنیم که قسمت‌های pad شده در روند آموزش مدل تاثیری نگذارند؛ که البته خود ابزار pytorch با استفاده از تابع

pack_padded_sequence بر روی ورودی لایه‌ی lstm به ما این اجازه را می‌دهد. در نتیجه‌ی این کار، loss نهایی ما تنها تابع کلمات اصلی جملات خواهندبود و نه کلمات pad شده. این روش نتیجه‌ی قابل قبول به ما داد که در بخش بعدی نتایج را خواهیم دید.

بعد از padding می‌توانیم داده‌ها را به مدل بدهیم. مدل ما یک لایه‌ی embedding دارد؛ که در واقع هر کلمه را با یک بردار مدل می‌کند؛ ورودی به pack_padded_sequence داده می‌شود تا به لایه‌ی lstm برود. خروجی لایه‌ی lstm هم به تعدادی لایه‌ی fully connected داده خواهد شد.

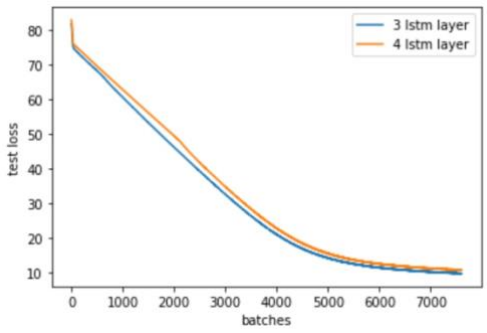
مقایسه‌ی مدل‌ها:

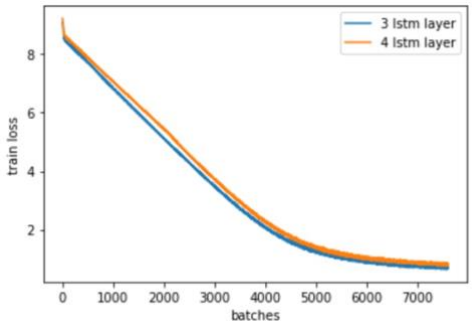
مدل اولیه‌ای که ما در نظرگرفتیم، یک مدل با ۳ لایه‌ی lstm و اندازه‌ی ۱۰۰ برای hidden، یک لایه‌ی fully connected و اندازه‌ی embedding برابر با ۵۰ است. سایر مدل‌ها با این مدل مقایسه شده و نتایج در زیر گزارش شده است.

- تعداد لایه‌های lstm:

در این بخش به جای۳ لایه‌ی lstm، از ۴ لایهه‌ی lstm استفاده کرده‌ایم:

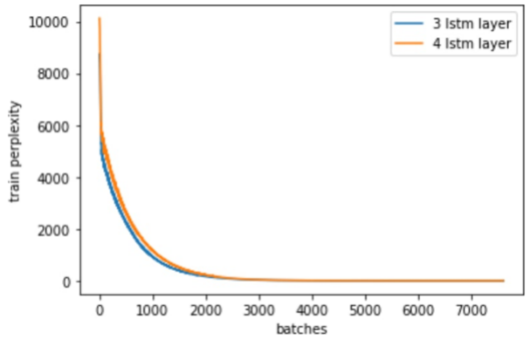
در اینجا تغییرات loss در طول آموزش را برای هر دو مدل، و بر روی هر دو دادگان train و test مشاهده می‌کنید:

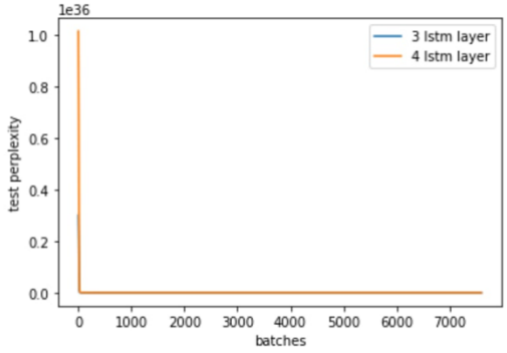




همانطور که از نمودارها مشخص است؛ مقدار loss تفاوت چندانی در این دو مدل ندارد. هر چند که در طول ۱۰۰ ایپاک، به نظر می‌رسد مدل با ۳ لایه‌ی lstm اندکی بهتر عمل کرده است و سرعت همگرایی هم تفاوتی ندارد. نقطه‌ی همگرایی در مدل با ۳ لایه، در loss کمتری است.

تغییرات perplexity برای دادگان آموزش و آزمون، نتایج زیر را ارائه کرده است:





باز هم به نظر می‌رسد که عملکرد ۳ لایه با اختلاف بسیار اندک، بهتر است؛ اما نهایتا به یک perplexity همگرا شده‌اند. از طرفی، عملکرد این دو مدل آنقدر نزدیک است که نقطه‌ی رسیدن به همگرایی هم تقریبا یکسان است. و سرعت تغییرات دو مدل هم تفاوت چندانی ندارد.

همچنین هیچ یک ازدو مدل باعث overfit نشده‌اند و خطای دادگان تست افزایش نیافته است.

جملات زیر توسط مدل با ۳ لایه تولید شده و کلمات داده شده به مدل: she could get rid of her car if بوده است که مدل ما تا ۲۰ کلمه بعد را پیش‌بینی کرده:

<s> she could get rid of her car if callous enduring lumps dabs scrambled curled water preferable dwarfed covered curled water ran ran cup grew outer passports cup consume

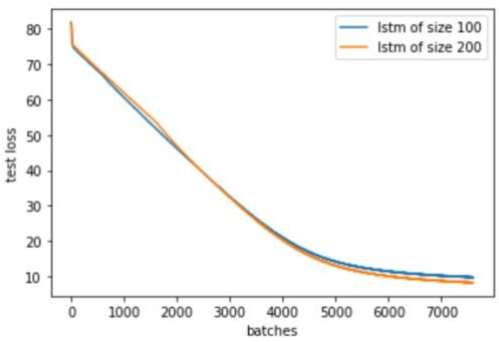
با همین کلمات ابتدایی، مدل ۴ لایه کلمات زیر را بدست آورده است:

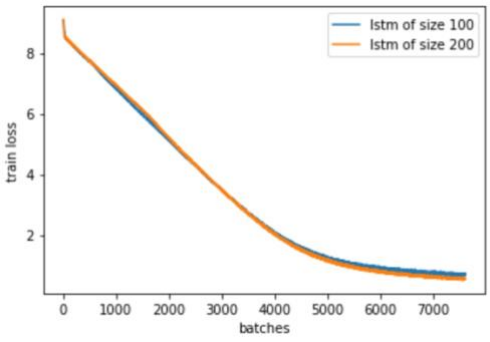
<s> she could get rid of her car if curled water dual worn callous handling handling disappointed opinions eager sensations salivating curled water water dragon worn cup hides handling

واقعا نمی‌توان گفت که کدام یک از این دو جمله واقعا عملکرد بهتری داشته‌اند! و به نظر می‌رسد که پارامترهای دیگری از این مدل نیاز به تغییر و اصلاح دارند.

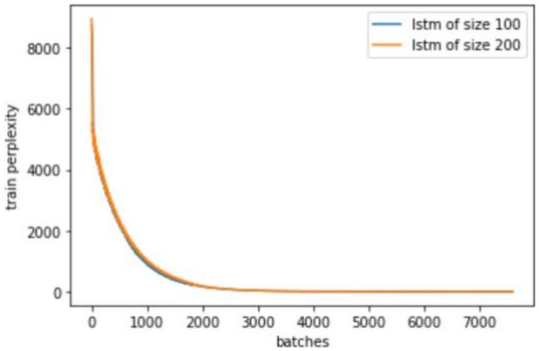
- اندازه‌ی hidden لایه‌ی lstm:

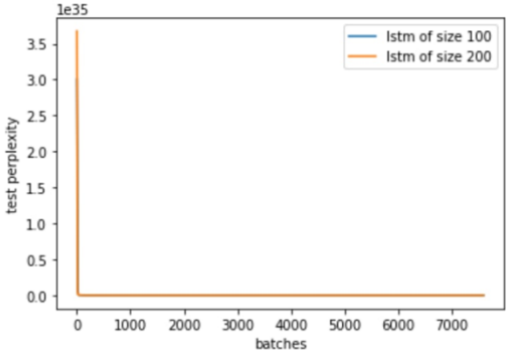
اندازه‌ی hidden را دو برابر، یعنی ۲۰۰ در نظر گرفتیم:





مدلی که اندازه‌ی hidden آن برابر ۲۰۰ در نظر گرفته شده‌است؛ نهایتا عملکرد بهتری در کم کردن خطای دادگان آزمون و آموزش داشته است. و طبعا از آنجایی که تعداد ایپاک‌های برابری





داشته‌ایم، سرعت این بهبود هم بیشتر بوده است. نقطه‌ی همگرایی برای مدل با اندازه‌ی ۲۰۰ هم در loss کمتری قرار دارد.

از روی نمودارها در رابطه با perplexity نمی‌توان اظهار نظر کرد اما با بسطدادن نتیجه‌ی lossها میتوانیم ادعای مشابهی برای این بخش داشته‌باشیم. و در اینجا هم هیچ یک از مدل‌ها دچار overfitting نشده‌اند.

جمله‌ی تولید شده با سید she could get rid of her car if در مدل با سایز ۱۰۰ همانطور که قبلا اشاره شده بود، در زیر آمده است:

<s> she could get rid of her car if callous enduring lumps dabs scrambled curled water preferable dwarfed covered curled water ran ran cup grew outer passports cup consume

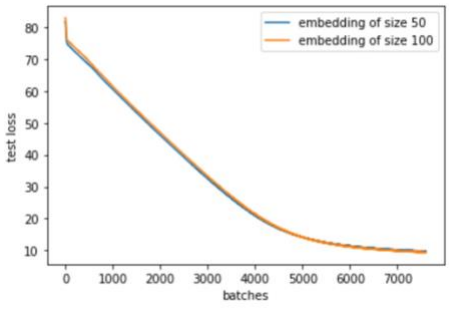
حالا جمله‌ای که با همان کلمات آغازین، در مدل با سایز lstm برابر ۲۰۰ تولید شده:

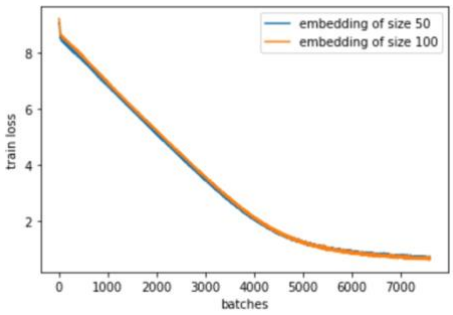
<s> she could get rid of her car if she could not remember the property of her life and her own life as she had gone

به وضوح جمله‌ی دوم به مراتب بهتر از جمله‌ی اول است و در نتیجه افزایش سایز لایه‌ی lstm منجر به نتیجه‌ی بسیار بهتری شده است. در نتیجه یکی از تغییراتی که مناسب است برای مدل اصلی در نظر گرفته شود اندازه‌ی لایه‌ی lstm می‌باشد.

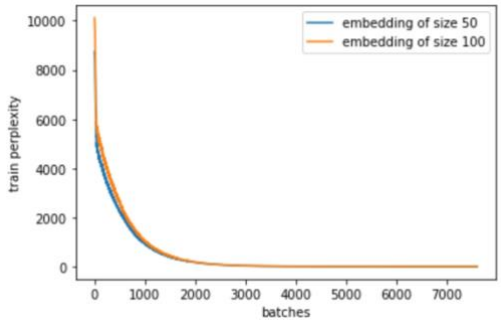
- اندازه‌ی embedding:

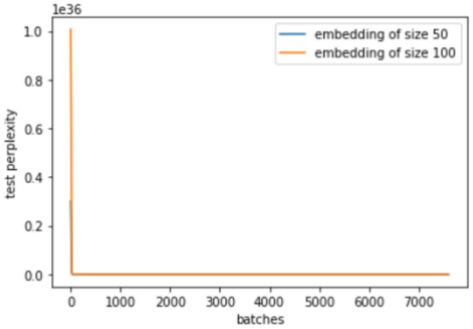
سایز خروجی لایه‌ی embedding را از ۵۰ به ۱۰۰ رساندیم:





این بار هم مدل تغییر چشمگیری نداشته اما در نقاط انتهایی، مدل با سایز embedding برابر ۱۰۰، خطای کمتری داشته. در نتیجه احتمالا embedding قوی‌تر تاثیر به‌سزایی در نتیجه‌ی مدل‌ها





خواهد داشت. اما سرعت کم شدن loss در دو مدل تفاوت چشمگیری ندارد و به نظر می‌رسد نقاط همگرایی تفاوتی ندارند!

همچنین perplexity در مدل با سایز embedding ۵۰ در اوایل بهتر است، اما سرعت همگرایی در سایز ۱۰۰ بیشتر است. در اینجا هم ovefit نداریم.

جمله‌ی تولید شده با سید she could get rid of her car if در مدل اصلی:

<s> she could get rid of her car if callous enduring lumps dabs scrambled curled water preferable dwarfed covered curled water ran ran cup grew outer passports cup consume

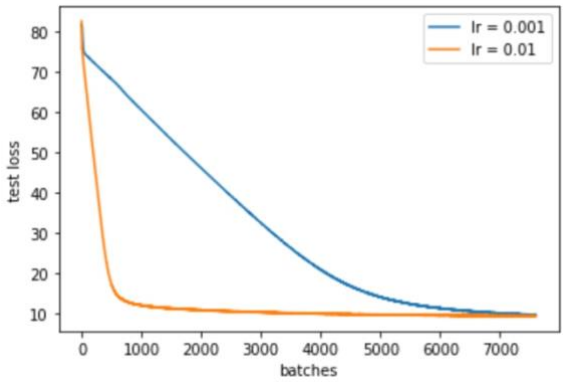
و جمله‌ای که با افزایش سایز embedding بدست آمده است:

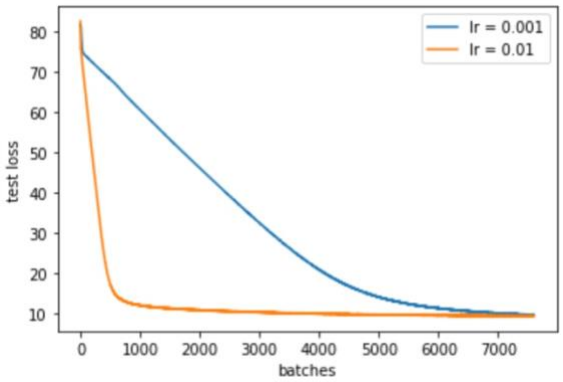
<s> she could get rid of her car if grab prick sensations plaster save angel worn water grab trunks save trunks water water water water water water water water

به نظر می‌رسد که این تغییر اصلا به نفع مدل ما نبوده و تعدادی کلمه در حال تکرار هستند. شاید یک نتیجه‌گیری این باشد که افزایش سایز لایه‌ی embedding باعث overfit بر روی کلمات (در بخش embedding) می‌شود.

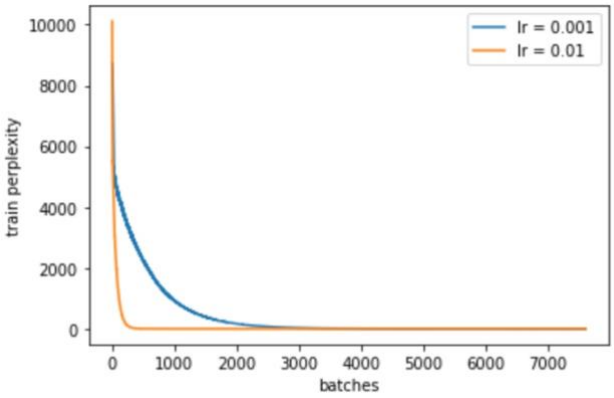
- تغییر در learning rate:

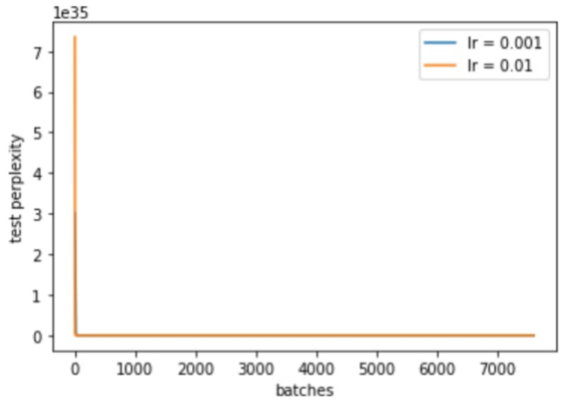
از نرخ یادگیر ۱۰ برابری استفاده کردیم. در مدل اصلی این عدد برابر ۰.۰۰۱ بود؛ که ما مدلی با نرخ یادگیری ۰.۰۱ را با آن مقایسه کردیم:





همانطور که انتظار می‌رفت، افزایش learning rate باعث افزایش سرعت یادگیری شده و مدل در تعداد اپیاک بسیار کمتری به دقتی که مدل اصلی در ۱۰۰ اپیاک رسیده، دست پیدا می‌کند. نقطه‌ی همگرایی به نظر می‌رسد که یکسان است و تعداد اپیاک بیشتر هم باعث overfitting در مدل نشده





است.

توقع داشتیم که perplexity هم با سرعت بیشتری همگرا شود که تصاویر نشان می‌دهد این اتفاق به درستی افتاده است.

جمله‌ی تولید شده توسط مدل اصلی:

<s> she could get rid of her car if callous enduring lumps dabs scrambled curled water preferable dwarfed covered curled water ran ran cup grew outer passports cup consume

و جمله‌ی تولید شده با نرخ یادگیری ۰.۰۱:

<s> she could get rid of her car if curled water curled water worn mouths contribution exceeded defied curled water grows grows wavelets grows exceeded wallet callous operator curled

به نظر نمی‌رسد که این افزایش سرعت همگرایی، نتیجه‌ی مطلوب‌تری داده باشد و همچنان جملات با معنی‌ای تولید نشده است.

حدود مقادیر loss مشاهده شده، قطعا به شیوه‌ی حل مشکل طول متفاوت کلمات بستگی دارد؛ چرا که زمانی که از روش padding استفاده می‌کنیم، در هر batch جملاتی با طول‌های متفاوت را به مدل نشان می‌دهیم اما در صورت استفاده از batchهایی با طول یکسان کلمات، می‌توانیم با ایجاد

بایاسی در سیستم باعث بروز اختلال در یادگیری شویم و این کار در مرحله‌ی اول خود را در مقادیر loss نشان خواهد داد.

همچنین پیش‌پردازش میتواند با کوچک و بزرگ کردن سایز vocab باعث تغییر در loss ما باشد. به این صورت که اگر ما پیش‌پردازش خیلی زیادی انجام دهیم و تعداد vocab را تا حد قابل توجهی کاهش دهیم؛ با کاهش تعداد کلاس‌ها و همچنین کاهش پیچیدگی داده‌های آموزش، مدل ما هم زودتر قادر به یادگیری خواهد بود و هم به خاطر تعداد کلاس کمتر، به loss کمتری دست پیدا خواهد کرد. اما این کار لزوما باعث افزایش دقت ما نخواهد شد؛ چون احتمالا بسیاری از جزئیات تاثیرگذار را هم در روند آموزش کم کرده‌ایم و عملا تفاوت‌ها را حذف کرده‌ایم.

5- قلم نویسنده

برای تولید جملات، کاری که صورت گرفته به این صورت است که تعدادی کلمه به عنوان کلمات شروع به مدل داده می‌شوند و مدل باید آنقدر کلمه تولید کند تا به انتهای یک جمله برسد.

مدل استفاده شده، یک مدل با سایز lstm ۲۰۰، اندازه‌ی embedding برابر ۱۰۰، ۳ لایه‌ی lstm و یک لایه‌ی fully connected است که بهترین مدل با توجه به نتایج قسمت قبل است.

برای انتخاب کلمات ما در هر مرحله ۳ کلمه با بیشترین احتمال را در نظر میگیریم؛ اگر یکی از آن‌ها <UKN> یا <pad> باشد؛ آن را حذف می‌کنیم و بعد کلمه با بیشترین احتمال را در نظر می‌گیریم.

در زیر می‌توانید جملات تولید شده توسط سه کتاب را ببینید که با کلمات ورودی به ترتیب:

Some times I want to

Here is where

So

There are

I love to

In the days of

و یک جمله‌ی خالی

تولید شده‌اند و تا زمانی که مدل به </s> برسد؛ ادامه یافته‌اند.

- کتاب zombie school:

<s> some times i want to be a zombie i said . </s>

<s> here is where it was a lot and the town 's mentor that it 's a human safe the town 's mentor had been a human . </s>

<s> so you have to be a zombie . </s>

<s> there are n't a zombie i said .</s>

<s> i love to the brain . </s>

<s> i like to run far and do you have to be a zombie i said .. </s>

<s> in the days of the stiffs were .</s>

<s> i did n't know that i had n't been a zombie i said . </s>

- کتاب a dream come true:

<s> some times i want to be of every rest and the term . </s>

<s> here is where the hell i 'd never been a few grin to humor the side in a </s>

<s> so i have a few moments in the side had still a </s>

<s> there are the news . </s>

<s> i love to the hell i 'd been in a life . </s>

<s> i like to run far a few days to a tangled glistening </s>

<s> in the days of a life in a few side . </s>

<s> i 'd been in a life in a few grin with the side had in the crop in a </s>

- کتاب a lady out of time:

<s> some times i want to see the future . </s>

<s> here is where that i . </s>

<s> so you 're going to be a few pounds in a few moments . </s>

<s> there are a son and the cloth out of a little voices of calf . </s>

<s> i love to know that but i would n't know what you 're here to me again . </s>

<s> i like to run far on the prostitute like and ... but that is the time hooker . </s>

<s> in the days of the night she 'd been shown in the plans . </s>

<s> i 'm a good opportunity to meet me . </s>

نشانه‌هایی که از context کتاب در جملات تولید شده مشاهده می‌شود، بسیار قابل توجه است.

** بنده مایل به دریافت گزارش متنی حاصل از بررسی گزارش هستم.

Sahar.rajabi76@gmail.com