N грами

Информацията за бъдещето не винаги е нещо добро.
Например коя дума най - вероятно би последвала?

Моля изчакайте …….

Повечето биха заключили, че една много вероятна дума е *на* , но също и *до, реда* . Формализирането на тази идея за предсказване на думи с вероятностни модели, се нарича N-грам модели, които предсказват следващата дума от предишните $N-1$ думи. Такива статистически модели на последователности от думи се наричат още езикови модели. Изчисляването на вероятността за следващата дума ще се окаже тясно свързано с изчисляването на вероятността за поредица от думи.

Следната последователност например има ненулева вероятност да се появи в текст:

... изведнъж забелязвам трима момчета, стоящи на тротоара.. докато същият набор от думи в различен ред има много малка вероятност:

... стоящи трима момчета забелязвам изведнъж, тротоара на...

N-грам модулите, които приписват условна вероятност на възможните следващи думи, могат да се използват за присвояване на обща вероятност на цяло изречение. Независимо дали се оценяват вероятностите на следващите думи или на цели поредици, моделът на N-грама е един от най-важните инструменти при обработката на реч и език.

N -грамите са от съществено значение във всяка задача, в която трябва да се идентифицират думи в шумен или двусмислен вход като **маркиране на част от речта** , **генериране на естествен език** и **сходство на думи** , както и в приложения от **идентификация на авторство** и **извличане на настроения** до системи за въвеждане на предсказуем текст за мобилни телефони.

. При **разпознаването на реч** , например, звуците на въведената реч са много объркващи и много думи звучат изключително сходно.

N-грам моделите също са от съществено значение в статистическия **машинен превод.**

В случая *на той информира репортерите за основното съдържание на…*

- *изявлението той информира репортерите за основното*

съдържание на

- *изявлението той информира репортерите за значимото съдържание на изявлението*
- ***той информира репортерите за основното съдържание на***

Посредством използване на N грам е по- вероятно да се изпише третото изречение, т.като е по-вероятно да има *информирани репортери* или *да се информират репортери,* отколкото някакъв друг случай, а *основното съдържание* е по-вероятно от *значимото съдържание* . Това позволява да се избере изречение с удебелен шрифт по-горе като изречение с най-плавен превод, т.е. това, което има най-голяма вероятност.

При **корекцията на правописа** трябва да се намери и коригират правописни грешки като онези, които случайно водят до истински думи:

Предсказването на думи също е важно за системите **за допълваща комуникация,** които помагат на хората с увреждания. Хората, които не могат да използват реч или жестомимичен език, за да общуват, като физика Стивън Хокинг, могат да общуват, като използват прости движения на тялото, за да избират думи от меню, които се изговарят от системата. Предсказването на думи може да се използва за предлагане на вероятни думи за менюто.

Преброяване на думи

Вероятността за поява на дадена дума в изречение се основава на преброяването (честотата на използването им).

Преброяването на естествен език се основава на елемент наречен корпус - онлайн колекция от текст или реч. Разгледани са два популярни корпуса Brown и Switchboard.

Първият е колекция от 1 милион думи от проби от 500 писмени текста от различни жанрове (вестници, романи, нехудожествена литература, академична литература и т.н.), събрани в университета Браун през 1963-64 г.

Колко думи има в едно изречение?

Зависи.

Ако не се броят препинателните знаци като думи ще едно число и друго ако се преброят. Пунктуацията е критична за намиране на граници на елементите (запетая, тире, точка, двоеточие) и за идентифициране на някои аспекти на значението (въпросителни знаци, удивителни знаци, кавички). За някои задачи, като маркиране на част от речта или синтактичен анализ или

синтез на реч, понякога препинателните знаци се третират като отделни думи.

Корпусът от телефонни разговори между непознати на Switchboard е събран в началото на 90-те години на миналия век и съдържа 2430 разговора средно по 6 минути, общо 240 часа реч и около 3 милиона думи. Такива корпуси от говорим език нямат пунктуация, но въвеждат други усложнения по отношение на дефинирането на думите.

Нека изречението от корпуса на Switchboard да бъде;

Ааа занимавам се главно с обработка нааа бизнес данни.

В този случай са налице т. нар паузи и междуметия, които могат да се игнорират или също да се отчитат при преброяването на думи според поставените цели.

Поради многозначността на думите и множеството на производните с общ корен е необходимо да се отчитат и другите значения. Това е прието да се нарича лема (lema). Речниците могат да помогнат при определянето на броя на лемите; речникови записи или извадки са много груба горна граница на броя на възможните леми (тъй като някои леми имат множество форми).

Речник на английските думи изброява 200 000 извадки. Ясно е, че колкото по-големи корпуси биват разглеждани, толкова повече видове думи се откриват.

За морфологично сложни езици като арабския задължително трябва да се използват N - грами .

Прости N -грами

С някои основни познания в теорията на вероятностите може да се запише следното:

вероятността за поява на дадена дума *w,* за някаква история *h*, или *P(w|h)*.

Нека да бъдат разгледани следните N - грами Предполага се, че историята *h* е „ *през есента е толкова красиво,…* “ и е необходимо да се знае вероятността следващата дума да е *P(през есента е толкова красиво, че).*

Как може да се изчисли тази вероятност? Един от начините е да се оцени от броя на относителните честоти. Например, може при много голям корпус, да се преброи колко пъти се появява *през есента е толкова красиво* и да се преброи колко пъти това е последвано от *че.* Това би отговорило на въпроса „Когато сме виждали историята h, колко пъти е била последвана от думата w“,

както следва:

$$\frac{P_{есен}}{P_{есен,че}} = P$$

С достатъчно голям корпус може да се изчислят тези употреби и оцени вероятността от уравнението.

Докато този метод за оценяване на вероятностите при директно преброяването работи добре в много случаи, се оказва, че дори мрежата не е достатъчно голяма, за да даде добри оценки в повечето случаи. Това е така, защото езикът е творчество; нови изречения се създават през цялото време и не винаги ще можем да преброим цели изречения. Дори простите разширения на примерното изречение може да имат нулев брой в мрежата (като „ *през есента в Балкана е толкова красиво, че* “).

По същия начин, ако е необходимо да се знае общата вероятност за цяла поредица от думи, като *през есента е толкова красиво,* би могло да се направи, като се попита „от всички възможни поредици от 5 думи, колко от тях са *през есента е толкова красиво*? ” Ще трябва да изчисли броя на *през, есента, е, толкова, красиво* , и раздели на сумата от броя на всички възможни последователности от 5 думи. Това е огромна изчислителна мощ и е трудно за оценка.

Поради тази причина трябва да се въведат по-умни начини за оценка на вероятността за дума *w,* за дадена история h, или вероятността за цяла последователност от думи *W* . Нека се започне с малко формализиране на нотацията. За представянето на вероятността конкретна случайна променлива X_i да приеме стойността „че“, или $P(X_i =$ „че“), или *P(че).*

Поредицата от *N* думи като $w_i ... w_n$ или „w”. За съвместната вероятност всяка дума в последователност да има определена стойност $P(X = w_1, Y = w_2, Z = w_3, ...,)$ ще се използва $P(w_i, w_2, ..., w_n)$.

Как могат да се изчислят вероятностите на цели последователности като $P(w_i, w_2, ..., w_n)$?

Чрез разлагане на вероятността,

$$\begin{aligned} P(X_1 \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1^2)\ldots P(X_n|X_1^{n-1}) \\ &= \prod_{k=1}^{n} P(X_k|X_1^{k-1}) \end{aligned}$$

Прилагайки това правило към думите, се получава:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\dots P(w_n|w_1^{n-1})$$
$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

Това показва връзката между изчисляването на съвместната вероятност на последователност и изчисляването на условната вероятност на дума, дадена на предишни думи.

Тези зависимости предполагат, че може да се оцени общата вероятност на цяла поредица от думи, като се умножи заедно с редица условни вероятности.

Сама по себе си тази нотация ще даде резултат сходен с предходния способ и не може просто да се изчисли и преброи колко пъти всяка дума се появява след всеки дълъг низ, защото езикът е креативен и всеки конкретен контекст може никога да не се е появявал преди

N - грам моделите вместо да изчисляват вероятността за дадена дума като се има предвид цялата й история, **приближават** историята само с последните няколко думи $P(\text{че}|\text{защото})$

Например **bigram** модела приближава вероятността за дадена дума всички предходни думи $P(w_n|w_n^{-1})$ като се използва само условната вероятност на предходната дума $P(w_n | w_{n-1})$. С други думи, вместо да се изчислява вероятността, то тя се приближава

Това предположение, че вероятността за една дума зависи само от предходната дума, се нарича предположение на **Марков** . Моделите на Марков са класът вероятностни модели, които предполагат, че може да се предвиди вероятността за някаква бъдеща единица, без да се интересува от момент t-2.

Може да се обобщи, че биграмата (следи до една дума в миналото) докато триграмата (три думи в миналото) и по този начин до **N-грамата** (която гледа N - 1 думи в миналото).

Така общото уравнение за това N-грамово приближение към условната вероятност за следващата дума в последователност е:

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$$

Като се има предвид предположението за биграмата за вероятността за отделна дума, може да се изчисли вероятността за пълна последователност от думи,

$$P(w_1^n) \approx \prod_{k=1} P(w_k|w_{k-1})$$

Оценка на тези вероятности може да се направи посредством максималната вероятност параметрите на N-грам модел, като се вземат преброявания от корпус и се нормализират , така че да варират между 0 и 1.

Показан е пример с изчисляване на вероятност на такъв модел.

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
```

Here are the calculations for some of the bigram probabilities from this corpus

$P(\texttt{I}\,|\,\texttt{<s>}) = \frac{2}{3} = .67$ $P(\texttt{Sam}|\,\texttt{<s>}) = \frac{1}{3} = .33$ $P(\texttt{am}|\,\texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>}|\,\texttt{Sam}) = \frac{1}{2} = 0.5$ $P(\texttt{Sam}|\,\texttt{am}) = \frac{1}{2} = .5$ $P(\texttt{do}|\,\texttt{I}) = \frac{1}{3} = .33$

For the general case of MLE N-gram parameter estimation:

Целта е да се оценява вероятността за N -грама чрез разделяне на наблюдаваната честота на определена последователност на наблюдаваната честота на префикс. Това съотношение се нарича **относителна честота** . По-горе бе казано, че това използване на относителни честоти като начин за оценка на вероятностите е пример за оценка на максималната вероятност. При оценката на максималната вероятност полученият набор от параметри максимизира вероятността за набора за обучение T , даден на модела M (т.е. $P(T|\text{M})$).

Предполагайки например, че думата *черноморски* се среща 500 пъти в корпус от милион думи. Каква е вероятността произволна дума, избрана от някакъв друг текст напр. един милион думи, да бъде думата *черноморски*?

Очевидно е 0,0005, но не е най-добрата възможна оценка на вероятността *черноморски* да се появи във всички ситуации; може да се окаже, че в някой

друг корпус или контекст *черноморски* е много малко вероятна дума.

N - грамният модел е добър пример за видовете статистически модели, които се наблюдават при обработката на реч и език. Вероятностите на *N* -грам модел идват от корпуса, върху който се обучава. Като цяло, параметрите на статистическия модел се обучават върху някакъв набор от данни и след това прилаган към модели и някои нови данни в някаква задача (като разпознаване на реч) и става ясно колко добре работят. Разбира се, тези нови данни или задача няма да бъдат точно същите данни, върху които е било извършено тренирането.

Може да се формализира идеята за обучение върху някои данни и тестване върху някои други данни, като се говори за тези два набора от данни като **обучителен набор** и **тестов набор** (или **корпус за обучение** и **тестов корпус**). По този начин, когато се използва статистически модел на езика, даден някакъв корпус от подходящи данни, тогава се започва с разделянето на данните на набори за обучение и тестове.

Тази парадигма за обучение и тестване може също да се използва за **оценка** на различни N-грам архитектури. Нека се сравният различни езикови модели (като тези, базирани на *N* -грами от различни порядъци N, или използвайки различни алгоритми **за изглаждане**. Възможно е да се вземе корпус и раздели на набор за обучение и тестов набор. Следва това обучаване на двата различни N-грам модела на набора за обучение и вижда кой моделира по-добре тестовия комплект. Но какво означава да „моделиране набора от тестове“? Има полезна метрика за това доколко даден статистически модел съвпада с тестова група, наречена **объркване**

Объркаността се основава на изчисляване на вероятността за всяко изречение в тестовия набор; интуитивно, който модел приписва по-висока вероятност на тестовия набор (следователно по-точно прогнозира тестовия набор) е по-добър модел.

Тъй като метриката за оценка се основава на вероятността на набора от тестове, важно е да не се допускат тестовите изречения в набора за обучение. Нека се изчисли вероятността за конкретно „тестово“ изречение. Ако то е част от учебния корпус, тогава погрешно ще му е присвоено изкуствено висока вероятност, когато се появи в тестовия набор. Тази ситуация се нарича **обучение върху тестовото множество** . Обучението върху набора от тестове

въвежда отклонение, което прави всички вероятности да изглеждат твърде високи и причинява огромни неточности.

В допълнение към наборите за обучение и тестове, други раздели на данни често са полезни и това са т. нар. **задържан** набор - допълнителен източник на данни, за да разширят обучителния набор.

Понякога трябва да има и множество набори от тестове. Това се случва, защото може да се използва определен тестов набор толкова често, че имплицитно да се настройва към неговите характеристики. Тогава определено ще има нужда от нов тестов комплект, който наистина е неизвестен. В такива случаи първоначалния тестов набор е набор **за разработка** или **devset**

Как се постига разделение на наборите за обучение, разработка и тестове? Има компромис, тъй като тестовия набор да бъде възможно най-голям и малък тестов набор може случайно да се окаже непредставителен. От друга страна, да съществуват възможно най-много данни за обучение. Като минимум е желателно да се избере най-малкия набор от тестове, който дава достатъчно статистическа мощност за измерване на статистически значима разлика между два потенциални модела. На практика често просто се разделят данните на 80% обучение, 10% разработка и 10% тест. Като се има предвид голям корпус, който да бъде разделен на тренировъчен и тестов, тестовите данни могат или да бъдат взети от някаква непрекъсната последователност от текст вътре в корпуса, или можем да премахнем по-малки „ивици“ текст от произволно избрани части на корпуса и да ги комбинира в тестов комплект.

Какви са възможните решения при изграждане на N-грам модели?

Като цяло трябва да е сигурно, че се използва обучителен корпус, който изглежда като тестов корпус. Особено не би бил избран за обучение и тестове от различни **жанрове** текст като вестникарски текст, ранна английска художествена литература, телефонни разговори и уеб страници. Понякога намирането на подходящ текст за обучение за конкретна нова задача може да бъде трудно ; за да изграждането на N-грам за предсказване на текст в SMS (услуга за кратки съобщения), има нужда от учебен корпус от SMS данни. За да изграждане на N-грам за бизнес срещи, ще трябва да има корпус от транскрибирани бизнес срещи.

4.5 Изглаждане

Възниква един основен проблем с процеса на оценка на максималната вероятност, който възникна в моделите за обучение на параметрите на N -грам модел. Това е проблемът с **оскъдните данни** , причинен от факта, че оценката на максимална вероятност се основава на определен набор от данни за обучение. За всяка N -грама, която се е появила достатъчен брой пъти, може да има добра оценка на нейната вероятност. Но тъй като всеки корпус е ограничен някои напълно приемливи последователности от думи със сигурност липсват в него. Тези липсващи данни означават, че матрицата на N грам модел за всеки даден обучителен корпус е обвързана да има много голям брой случаи на предполагаема „нулева вероятност.

Освен това, методът за оценка на максималната вероятност също дава лоши оценки, когато броят е различен от нула, но все още е малък.

Необходим е метод, който може да помогне да се получат по-добри оценки за тези нулеви или нискочестотни преброявания. Оказва се, че нулевото отчитане причинява и друг огромен проблем. Това е т.нар. показател **за объркване** изисква изчисляване на вероятността за всяко тестово изречение. Но ако тестовото изречение има N -грам, която никога не се е появявала в набора за обучение, оценката на максималната вероятност на вероятността за тази *N -грама*, а оттам и за цялото тестово изречение, ще бъде нула. Това означава, че за да се оценят собствените езикови модели, трябва да се модифицира метода за максимална оценка, за да се присвои някаква различна от нула вероятност на всяка *N -грама*, дори такава, която никога не е била наблюдавана в обучението.

За решаване на този проблем е въведен термина **изглаждане** на такива модификации, които се отнасят до лошите оценки, които се дължат на променливост в малки набори от данни посредством отчитане малко напред ще се намали малко вероятностна маса от по-високите стойности и вместо това ще се натрупа върху нулевите стойности, правейки разпределението малко по изгладено (по-малко назъбено).

Изглаждане на Лаплас

Един прост начин за изглаждане може да бъде просто да се вземе матрица от преброявания на биграми, преди да са нормализирани във вероятности, и да се добави едно към всички преброявания. Този алгоритъм се нарича **изглаждане на Лаплас** или закон на Лаплас

Неизгладената оценка на максималната вероятност на вероятността за дума W_i е нейният брой c, нормализиран от общия брой токени на дума N:

$$P(w_i) = \frac{c_i}{N}$$

Изглаждането на Лаплас просто добавя единица към всеки брой, при положение, че общият брой на всички думи в един корпус е V *думи и всяка от тях е* увеличена, също трябва да се коригира и знаменателя, за да се вземат предвид допълнителните V наблюдения

$$P_{\text{Laplace}}(w_i) = \frac{c_i + 1}{N + V}$$

Вместо да се променят както числителят, така и знаменателят, е удобно да се опише как алгоритъмът за изглаждане влияе на числителя, като се дефинира коригиран брой c^*. Този коригиран брой е по-лесен за директно сравняване с броя на оценката за поява и може да бъде превърнат във вероятност чрез нормализиране с N. За да дефиниране на този брой се променя само числителя, в допълнение към добавянето на един също ще трябва да се умножи по коефициент на не анализиране: $\frac{N}{N+V}$

Така се получава

$$c_i^* = (c_i + 1)\frac{N}{N+V}$$

Сега може c* да се превърне във вероятност p* чрез нормализиране с N.

Начин за разглеждане на изглаждането е като понижаване на някои ненулеви преброявания, за да се получи вероятностната маса, която ще бъде присвоена на нулевите преброявания. Така вместо да се позовава на намалените стойности c*, може да се състави алгоритъм за изглаждане по отношение на относителна отстъпка d_c, съотношението на намалените стойности към първоначалните стойности:

$$d_c = \frac{c^*}{c}$$

За преброяване на изгладени биграми с добавяне на единица трябва да се увеличава броя на общите типове думи в речника *V*.

За изграждане на езикови модели са разработени множество инструментариум, но два са често използвани SRILM и Cambridge-CMU. И двата са публично достъпни и имат сходна функционалност.

Заключение

Вероятността за *N -грам* е условната вероятност за поява на дума, въз основа на предишните *N -1* думи. Вероятностите могат да бъдат изчислени чрез просто преброяване в корпус и нормализиране (оценка на максималната вероятност) или могат да бъдат изчислени чрез по-сложни алгоритми. Предимството на *N -грама* е, че се възползва от много богати лексикални познания. Недостатък за някои цели е, че те са много зависими от корпуса, на който са били обучени.

Алгоритмите за изглаждане предоставят по-добър начин за оценка на вероятността за *N -грам* от оценката на максималната вероятност. Често използваните алгоритми за изглаждане разчита на броя *на N -грам* от по-нисък ред чрез отстъпка или интерполация.