ВЫВОД И ОЦЕНКА ПАРАМЕТРОВ ДАЛЬНОДЕЙСТВУЮЩЕЙ ТРИГРАММНОЙ МОДЕЛИ ЯЗЫКА

INFERENCE AND ESTIMATION OF A LONG-RANGE TRIGRAM MODEL

Протасов С.В. *(svp@tj.ru)*
Московский физико-технический институт (Государственный университет)

В докладе описывается простая вероятностная грамматика связей (Link Grammar), известная также, как “Модель дальнодействующих триграмм” (Long-range Trigram Model). Эта вероятностная модель языка расширяет триграммные модели, предсказывая слова не только по двум непосредственно предшествующим словам в предложении, но и потенциально по любой паре стоящих рядом слов, которые лежат внутри этого же предложения. Таким образом, триграммная модель может пропускать менее информативные слова для более точного пргноза. Лежащая в основе “грамматика” есть не более, чем множество пар слов, которые могут быть связаны вместе через несколько разделяющих слов; это множество слов получается автоматически из корпуса текста, используемого для “обучения модели” грамматики. В докладе представлены результаты экспериментов, совершенные на корпусе предложений русского языка.

1. Вступление

В данной работе мы исследуем модель языка, которая может использоваться для практических задач, где требуется вероятностная оценка корректности предложений. В работе [Protasov 06]автором исследовалась более сложная модель и её реализация не позволила провести обучение (тренировку) на большом корпусе. В частности, использовался корпус около 3 тыс предложений со словарём примерно 300 слов. Конечно же в реальных задачах нам потребуются модели, которые позволяют обрабатывать корпуса размером на несколько порядков больше. Далее мы будем обсуждать одну из таких моделей, которая, по сути, является упрощением контекстно-свободной модели языка из работы [Protasov 06].

Наиболее широко используемой статистической моделью языка в настоящий момент является так называемая *триграммная модель*. В этой простой модели слово предсказывается на основе только лишь двух слов, непосредственно стоящих перед ним. Простота *триграммной модели* одновременно является и ее наибольшим преимуществом, и недостатком. Преимущество модели заключается в том, что для оценки параметров модели языка существует достаточно простой и быстро работающий алгоритм, который может обработать сотни миллионов слов текста. Реализация модели будет содержать внутри всего лишь поиск по большой таблице, что достаточно просто в практическом плане. Все новые статистические модели практически всегда оцениваются по отношению к триграммной модели. На сегодняшний день многие успешные системы распознавания речи в той или иной форме используют именно n-граммную модель (где n=2,3) [Jelinek, 97]. Несмотря на свои успехи триграммная модель ничего не знает о богатых синтаксических и семантических связях, которые содержат естественные языки, позволяя им быть легко распознаваемыми и понимаемыми людьми. Во многих реальных предложениях зависимые слова находятся на довольно большом расстоянии в 5-7 слов и триграммная модель никак не может учесть эти связи. Использование n-граммных моделей с n=5,6,7 требует гиганских ресурсов и сталкивается с проблемой “редких данных”.

Вероятностная грамматика связей была предложена как подход, который сохраняет достоинства и вычислительные преимущества триграммной модели, и в то же время включает дальнодействующие зависимости и более сложную информацию в статистическую модель [Lafferty et al. 92]. В этом докладе будет представлена реализация очень простого варианта *вероятностной грамматики связей,* которая (реализация) применима для любого естественного языка, включая русский. Грамматика расширяет *триграммные модели* через разрешение связей между словами, предшествующими не только в пределах двух предыдущих слов, но и потенциально находящимися на большем расстоянии от предсказываемого слова в пределах предложения. Таким образом *дальнодействующая триграммная модель* может пропускать малоинформативные слова и улучшать предсказуемость в модели. Лежащая в основе грамматика представляет собой множество пар слов, которые могут быть соединены друг с другом через несколько промежуточных слов. Впервые *дальнодействующая триграммная модель* была предложена в работе [Pietra et al. 94], где она исследовалась на англоязычном материале, но к сожалению резуль-

таты исследований ученых из IBM не были подтверждены независимо, не говоря уже о доступности каких-либо программных реализаций модели, а публикации по исследованию на корпусах других языков отсутствуют до сих пор.

Далее во втором разделе будет кратко описано введение в *дальнодействующую триграммную модель* и показано, как она может быть представлена в виде *вероятностной грамматики связи*. Грамматика парных слов автоматически выводится из корпуса обучающего текста. Хотя взаимная информация слов также может использоваться для эвристического вывода парных слов, сам по себе этот подход не приносит адекватных результатов. В третьем разделе будет описан алгоритм, адаптирующий критерий *взаимной информации* для наших целей. В последнем разделе представлены результаты экспериментов, совершенных на русскоязычном материале.

2. Дальнодействующая триграммная модель

В качестве примера рассмотрим рисунок 1. На диаграмме представлена *связка* (linkage) предложения “Если у Вас есть ... заработать.”, согласно формализму, впервые введенному в [Sleator and Temperley, 91], важными свойствами связки является непересечение связей, их связность (отсутствие неприсоединенных областей), единственность связей (каждая пара слов соединена только одной связью). Рассматривая вероятностную модель, мы считаем, что каждое слово генерируется из биграммы заканчивающейся словом, примыкающим к генерируемому слову слева. Таким образом, первая правая скобка сгенерирована на основе биграммы (сайт|(), а первое слово “сайт” сгенерировано из биграммы (есть|свой). Слово “то” сгенерировано из биграммы (⊥,|Если), где ⊥ является специальным словом-границей.

Рис. 1. Дальнодействующие триграммы

Для описания модели более детально, рассмотрим следующее описание стандартной триграммной модели. Модель может быть рассмотрена как простой конечный автомат, генерирующий предложения. Состояния этого автомата проиндексированы парами слов. Добавив слово-границу ⊥ в наш словарь слов, мы зададим начальное состояние конечного автомата как (⊥,⊥). Когда автомат находится в каком-либо состоянии (w_1,w_2), он может перейти в состояние (w_2,w_3), с вероятностью $t(w_3|w_1,w_2)$ и остановится с вероятностью $t(\perp|w_1,w_2)$, таким образом остановив предложение.

Наша расширенная триграммная модель может быть описана похожим образом. Для ссылки на состояния автомата используются пары слов, но состояние $s = (w_1,w_2)$ теперь может быть одним из трех: останов (halt), шаг (step), ветвление (branch) с вероятностями *d(halt|s), d(step|s), d(branch|s)* соответственно. В случае выбора состояния *step* или *branch*, следующее слово *w* генерируется с триграммной вероятностью $t(w|w_1,w_2)$. Но в случае выбора *branch* генерируется дополнительное слово *w′* на основе дальнодействующей триграммы $l(w'|w_1,w_2)$. Например, в процессе генерирования связки из примера выше, состояние с индексом *s* = (то,регистрация) приводит к состоянию step с вероятностью *d(step|s)* и слово “позволит” затем генерируется с вероятностью *t*(позволит|то,регистрация) С другой стороны, состояние s = (⊥,Если) ответвляется с вероятностью *d(branch|s)* и затем из этого состояния генерируется слово “у” и слово “то” с вероятностью *t*(у| ⊥Если) и *l*(то| ⊥Если).

В результате все слова в связках, как на примере выше, имеют ровно одну связь слева и ноль, одну или две связи справа. Если мы пронумеруем слова в предложении *S* от 1 до |*S*|, тогда вполне удобно обозначать через $\lhd i$ индекс слова, которое генерирует слово слева от i-го в предложении. Таким образом, *i* соединено слева с $\lhd i$. Например, на связке из примера выше мы видим, что $\lhd 9 = 8$, $\lhd 8 = 1$, и $\lhd 26 = 18$. Подобная запись позволяет нам записать вероятность предложения как $P(S) = \Sigma_{L(S)} P(S,L)$, где *L(S)* есть набор всех связок *S* и где соединяющая вероятность *P(S,L)* расписывается как

$$P(S,L) = \prod_{i=1}^{|S|} d(d_i|w_i,w_{i-1})t(w_i|w_{i-2}w_{i-1})^{\delta(i-1,\lhd i)}l(w_i|w_{\lhd i-1}w_{\lhd i})^{1-\delta(i-1,\lhd i)} \quad (1)$$

Здесь $d_i \in halt, step, branch, (i, j)$ равен единице, если $i = j$, и нулю, если не равен. Индекс $<i$ должен пониматься по отношению к заданной связке L.

В терминах *грамматики связей* [Sleator and Temperley, 91] переменные *halt, step* и *branch* эквивалентны трем простым *дизъюнктам*, определяющим, как заданное слово соединяется с другими словами. Значение *halt* соответствует дизъюнкту, имеющему один левый коннектор (без метки) и не имеющий правых коннекторов. Значение *step* соответствует дизъюнкту, имеющему единственный левый и единственный правый коннектор. Значение *branch* соответствует дизъюнкту имеющему один левый коннектор и два правых коннектора. В формализме данной грамматики вероятностная модель (1) является простым вариантом более общей вероятностной грамматики связей, представленной в работе [Lafferty et al. 92].

На этом мы закончим сверхкраткое введение в дальнодействующие триграммные модели и за дополнительной информацией рекомендуем обратиться к работе [Pietra et al. 94]. Там же дано описание эффективного алгоритма “обучения” модели (что равносильно выводу грамматики). Целью алгоритма является увеличение суммы (1) по всем предложениям в обучающем корпусе. Алгоритм “обучения” хоть и является разновидностью EM (Expectation-maximization, разновидность алгоритма максимизации правдоподобия) [Baum 72], в действительности довольно сильно отличается от популярного подхода Inside-Outside [Lari and Young, 90], который часто используется для обучения формальных вероятностных моделей [Manning and Shutze, 99].

3. Вывод грамматики

Вероятностная модель (1), описанная в предыдущем разделе, делает свои предсказания на основе как обычных триграммных моделей, так и на основе дальнодействующих триграмм. Мы можем разрешить использовать связи со словами, присоединяющееся слева к любому слову. Это соответствует “грамматике”, которая разрешает дальнодействующие связи между любыми двумя словами. Число возможных *связок* для такой грамматики растет очень быстро с увеличением длины предложения: если предложение состоящие из 10 слов имеет всего-лишь 835 *связок*, то предложение, состоящее из 25 слов уже имеет 3 192 727 797 *связок*. Однако большинство дальнодействующих связей в этих связках скорее всего будут неправильными. Получившаяся вероятностная модель имеет слишком много параметров, которые не могут быть достаточно точно оценены. А для целей качественного обучения нам требуется высокое отношение “число примеров/число параметров”.

L	R	$log(Gain_{LR})$	$d(branch_{LR}\|L)$	$d_{LR}(halt)^{-1}$
(	)	11.05	0.8558	4.4
Если	то	8.81	0.3541	7.1
либо	либо	8.44	0.3398	4.2
"	"	8.33	0.2171	7.9
Ни	ни	7.92	0.4228	2.6
Чем	тем	7.78	0.4414	5.0
столько	сколько	7.76	0.2661	2.6
Чем_больше	тем	7.66	0.9585	4.2
Что_касается	то	7.47	0.7549	3.5
ни	ни	7.43	0.2123	2.8
Чем_больше	тем	7.66	0.9585	4.2
Ни_одна	не	5.92	0.9364	2.8
Чем_дольше	тем	5.83	0.9157	4.2
кроме_тех_случаев	когда	5.14	0.9241	1.2
только_в_том_случае	если	7.21	0.7349	1.1
Никакой	не	5.22	0.7549	3.1
Что_касается	то	7.47	0.7549	3.5
Интересно	?	5.73	0.2437	8.8
Даже_если	все_равно	5.25	0.1187	8.7
Во-первых	во-вторых	6.08	0.1294	8.5
Неужели	?	7.17	0.4026	7.6
Разве	?	6.57	0.2864	7.6
Ах	!	6.54	0.4918	7.4
Если	то	8.81	0.3541	7.1
Почему	?	7.41	0.3296	7.0
Сначала	потом	5.77	0.2168	6.3
Одна	другая	5.04	0.0877	6.1

L	R	$log(Gain_{LR})$	$d(branch_{LR}\|L)$	$d_{LR}(halt)^{-1}$
отличить	от	6.49	0.6775	1.7
не_обращал	внимания	6.12	0.5178	2.1
избавить	от	5.88	0.7158	1.2
споткнулся	упал	5.86	0.3114	2.3
Одним_из	является	5.81	0.4638	4.3
отделить	от	5.68	0.6820	1.8
превратить	в	5.36	0.6560	1.7
Делать	нечего	5.30	0.4220	2.1
прижала	к	5.17	0.6869	1.3
обращать	внимание	4.91	0.2997	2.0
Целью	является	4.84	0.5089	2.1
нашелся	ответить	4.55	0.2091	1.9
Прошло	прежде_чем	4.53	0.1653	3.0
поблагодарить	за	4.48	0.4136	1.5

Таблица 1. Примеры пар слов

Раз неограниченная грамматика непрактична, мы попробуем ограничить грамматику через разрешение только тех дальнодействующих связей, которые приносят наибольшие улучшения в вероятностную модель. В идеале нам нужно автоматически выявлять пары слов, такие как “(” и “)” с дальнодействующими корреляциями, которые могут быть хорошими кандидатами на соединение через дальнодействующую связь. Мы можем поискать такие пары через просмотр слов с высокой взаимной информацией. Но если мы представим, что мы уже включили связи всех ближайших соседей в нашу модель, как в случае модели (1), то у нас не будет точек для связывания слов *L* и *R*, независимо от того, насколько велика их взаимная информация, ведь слово *R* уже хорошо предсказывается непосредственными предшественниками. Вместо этого мы будем искать связи между словами, которые имеют потенциал улучшения модели только по сравнению с обычными короткими связями.

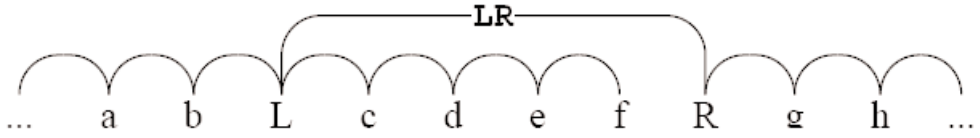

Рис. 2. Модель LR

Для нахождения таких пар используем следующий подход. Пусть V словарь языка. Для каждой пары (L,R) 2 V × V сконструируем модель PLR, которая содержит все связи биграммами с одной дополнительной дальнодействующей связью, идущей от L к R. На основе анализа корпуса русскоязычных предложений мы определим пользу пары (L,R) по сравнению с биграммной моделью. Мы выбрали модель PLR достаточно простой, чтобы параметры всех |V |2 возможных моделей оценивались параллельно. Затем мы отсортируем модели согласно их правдоподобию GainLR [Pietra et al. 94], которую каждая модель показывает на обучающем корпусе, и выберем те пары (L,R), которые соответствуют самым лучшим моделям. Этот список пар слов и будет составлять нашу новую “грамматику”, описанную в предыдущем разделе.

4. Результаты экспериментов

Этот раздел представляет результаты обучения наших дальнодействующих триграммных моделей на корпусе предложений, собранных через интернет. Наш обучающий корпус состоял из более чем 11 млн. предложений, содержащих примерно 150 млн. слов. Таблица 1 включает примеры пар слов, которые были получены после использования формул из раздела 3. Напомним, что эти пары были получены при первом шаге обучения грамматики связей, которая позволяет дальние связи между одной фиксированной парой слов. Каждая пара проверяется уменьшением энтропии, которая ее односвязная модель достигает по сравнению с биграммной моделью. В таблице это улучшение показано в 3-м столбце. Мы сразу отсекаем все пары, которые не приводят к уменьшению энтропии. В первой секции таблица содержит пары, которые приводят к наибольшему уменьшению энтропии. Четверый столбец таблицы дает значения вероятности $d(branch_{LR}|L)$. Это значение показывает вероятность, с которой *L* генерирует *R* с некоторого расстояния в соответствии с обучаемой моделью. Вторая секция таблицы

включает примеры пар с высоким значением вероятности $d(branch_{LR}|L)$. Пятый столбец таблицы дает значения вероятности $d_{LR}(halt)^{-1}$. Поскольку в обучающих данных число слов между *L* и *R* убывает геометрически со средним $d_{LR}(halt)^{-1}$, то большое значение в этом столбце указывает, что *L* и *R* находятся в среднем на достаточно большом расстоянии. Третья секция таблицы приводит примеры таких пар. В заключение, четвертая секция таблицы показывает пары, где одно из слов является глаголом, и только некоторые из этих пар с наибольшим уменьшением энтропии показаны в таблице.

Мы довольны результатами, так как полученные списки пар практически не содержат мусора, который в избытке появляется при использовании других методов. Однако из-за нечеткости критерия “что есть мусор”, нам очень сложно провести численное сравнение. Более того, нам вообще хотелось бы избежать человеческой оценки качества связей и использовать более формальные оценки.

5. Выводы и планы

Полученные данные позволяют сделать вывод, что модель дальнодействующих триграмм представляет собой еще один инструмент корпусной лингвистики. Этот инструмент, в частности, позволяет автоматически устанавливать факт наличия синтаксической связи между словами, не стоящими рядом. Полученная “грамматика пар слов” может быть использована для инициализации более сложных вероятностных моделей. Исследование пар, отфильтрованных по частям речи, может помочь в изучении “дальних” валентностей глаголов, а также составлению списка глаголов, потенциально имеющих большое число валентностей. Было бы интересно изучить таким способом какой-либо мертвый язык, имеющий достаточно большой корпус текстов. Однако наша долгосрочная цель не словарь парных слов, а более мощная статистическая модель языка. Если в процессе тренировки модели мы получаем качественный словарь, содержащий мало мусора, то это хорошее свидетельство того, что мы движемся в правильном направлении. После того как мы получили список пар кандидатов, имеющих дальние связи, нам нужно провести несколько шагов пере-инициализации параметров Expectation Maximization. Данная процедура может существенно изменить вероятности связей и даже сделать какие-либо из них несущественными для грамматики. Несколько шагов тренировки могут привести к дальнейшему отсеву мусора среди пар кандидатов. К сожалению, делать переобучение нужно не целиком в основной памяти компьютера (для больших словарей порядка 100 тыс слов её может не хватить), а через последовательную обработку файлов корпуса на жестком диске. Данный этап работы автором еще не завершен. Кроме этого, качественная статистическая модель обязательно содержит процедуры сглаживания, а это требует дополнительного программирования. Так как каждая пара приводит к уменьшению кросс-энтропии корпуса, то все пары в сумме также гарантировано должны приводить к снижению кросс-энтропии. Однако нам неизвестно, насколько велико будет суммарное улучшение и будет ли оно существенно лучше n-gram моделей. Проводить сравнение несглаженной дальнодействующей модели со сглаженной n-gram моделью не вполне корректно, так как несглаженные модели существенно хуже, чем сглаженные. После настройки параметров вероятностной модели, мы можем подключить нашу модель языка в какую-либо практическую систему для измерения качественных результатов. К примеру известно, что в системах распознавания речи, где также используются статистические модели языка, число ошибок линейно уменьшается в зависимости от кросс-энтропии. Мы также можем подключить модель языка к системе статистического машинного перевода и измерить улучшение по стандартной BLEU метрике, хотя у нас есть подозрения, что BLEU метрика не увидит улучшения, так как использует ngram-совпадения при сравнении переводов. Человеческие оценки качества перевода несколько затратны и не могут быть осуществлены для больших корпусов. Таким образом, за неимением лучшего, мы будем использовать кросс-энтропию на тестовом корпусе как самый главный критерий качества нашей языковой модели.

Список литературы

1. [Protasov 06] Протасов С . В . Обучение с нуля грамматики связей русского языка. //Десятая национальная конференция по искусственному интеллекту с международным участием., КИИ-2006.

2. [Lafferty et al. 92] Lafferty J. Sleator D. Temperley D. Grammatical Trigrams: A Probabilistic Model of Link Grammar. //Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language, 1992.

3. [Pietra et al. 94] Pietra S., Pietra D., Gillet J., Lafferty J., Prinz H., Ures L. Inference and Estimation of a Long-Range Trigram Model. //Grammatical Inference and Applications, Second International Colloquium, ICGI-94, 1994.

4. [Sleator and Temperley, 91] Sleator D. Temperley D. Parsing English with a Link Grammar.//Carnegie Mellon University Computer Science technical report CMU-CS-91-196, 1991.

5. [Jelinek, 97] Jelinek F. Statistical Methods for Speech Recognition. //MIT Press. ISBN: 0-262-10066-5. M.: 1997.

6. [Manning and Shutze, 99] Manning C., Schutze H. Foundations of Statistical Natual Language Processing. //Cambridge, MA: MIT Press.M.: 1999.

7. [Baum 72] Baum L. E. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. //Inequalities, 627(3):1-8,M.: 1972.

8. [Brown, 92] Brown P. F. Stepen A. L. An estimate of an upper bound for the entropy of English. //Computational Linguistics. 1992.

9. [Lari and Young, 90] Lari K. Young S. J. The estimation of stochastic context-free grammars using the inside-outside algorithm. //Computer Speech and Language. 1990.