Параметрическая оптимизация точности морфологической разметки текстов

Рысаков С.В., НИУ Высшая школа экономики, МИЭМ
srysakov@hse.ru
Клышинский Э.С., ИПМ им. М.В. Келдыша РАН
klyshinsky@mail.ru

Аннотация

Статья знакомит читателя с базовыми понятиями параметрической оптимизации. Описывается разработанная модель аппроксимация вероятности, функции-счётчики и коэффициенты корреляции. Небольшое внимание уделено методу полного перебора, в результате работы которого достигнуты новые показатели точности. В конце приведена модификация метода снятия омонимии, разработанная авторами.

1 Введение

В предыдущей статье [Рысаков, Клышинский, 2015] проводилось сравнение различных подходов к решению задачи снятия омонимии. Методы, отличившиеся наилучшими показателями точности, объединяло одно свойство: в их основе лежали скрытые марковские модели, широко используемые в компьютерной лингвистике графические модели, имеющие скрытые и наблюдаемые переменные (состояния). Наблюдаемым переменным в рамках предыдущей статьи соответствовали слова в предложении и знаки пунктуации, а скрытым – теги морфологической разметки. В общем случае, для поиска наиболее вероятной последовательности тегов требовалось максимизировать произведение частных вероятностей:

$$T = \underset{t \in S}{\operatorname{argmax}} \prod_{i=1}^{N} P(t_i | t_{i-1}, Q)$$

где N – количество слов в предложении, S – множество всех последовательностей (цепочек) тегов длины N, Q – множество наблюдаемых переменных, содержащее слова и знаки пунктуации ($Q = \{w, p\}$).

Очевидно, что выбор омонима зависит от функции вероятности P и её параметров, используемого множества наблюдаемых переменных Q. В данной статье мы покажем результаты выбора параметров данной модели.

2 Выбор параметров модели

Предположим, что у нас отсутствует формула, рассчитывающая вероятность следования тега t_i за тегом t_{i-1}, но зато имеется внушительный корпус текстов с морфологической разметкой. В этом случае можно рассчитать приблизительное значение этой вероятности на основе статистических данных с помощью формулы условной вероятности:

$$P(t_i | t_{i-1}) \approx \frac{C_2(t_{i-1}, t_i)}{C_1(t_{i-1})}, \qquad (1)$$

где C – количество последовательностей тегов (C_1 – униграмм и C_2 – биграмм), встреченных в корпусе. При увеличении корпуса распределение вероятностей постепенно стабилизируется. Если же конкретный набор данных нам не встретился, для функции C будет использоваться сглаживание Лапласа, то есть минимальным возвращаемым значением будет единица, а не ноль.

2.1 Выбор параметров функции

Статистика учитывает встречаемость не только тегов, но и наблюдаемых параметров и их сочетаний. К примеру, $C_p(t_{i-1}, p_{i-1}, t_i)$ соответствует числу последовательно идущих тегов t_{i-1} и t_i, разделённых *знаком пунктуации* p_{i-1}, а $C_w(w_i, t_i)$ показывает количество встреченных *начальных форм слов* w_i, имеющих тег t_i. К числу параметров был добавлен и так называемый *класс неоднозначности* a_w [Cutting et al., 1992] – множество возможных тегов для данного слова w. Хотя класс неоднозначности сам по себе не является наблюдаемой переменной, его легко узнать, пользуясь морфологическим словарём.

Определим множество переменных модели как $Q = \{w, p, a_w\}$. На основе этих переменных выберем аппроксимирующую функцию.

2.2 Выбор функции аппроксимации

Перепишем формулу (1) с учетом множества Q в следующем виде:

$$P(t_i | t_{i-1}) \approx C_2(t_{i-1}, t_i, Q) * C_1^{-1}(t_{i-1}, Q),$$

или в более общем виде:

$$P(t_i|t_{i-1},p_{i-1},Q) \approx \prod_{k\in I} C_k(t_{i-1},t_i,Q)^{r_k} \quad (2)$$

где C_k – k-е статистическое значение, рассчитанное из корпуса, I – множество индексов статистических значений, а r_k – соответствующий степенной показатель. Помимо значений C_1 и C_2 мы также можем использовать C_p, C_w, а также другие значения. Для перечисленных выше значений I={1, 2, p, w}, но в дальнейшем мы будем использовать более широкое множество. Степенной показатель также может отличаться от значений 1 и -1.

Таким образом, вместо формулы условной вероятности мы будем использовать формулу, учитывающую большее количество параметров с различными степенными показателями, принимающую в качестве аргумента различные параметры модели. Формально полученные цифры не являются вероятностями (до тех пор, пока не будет доказано обратное).

Для фиксированного размеченного корпуса мы можем подобрать значения степенных показателей так, чтобы снятие омонимии с использованием полученной формулы давало максимальную точность. Данная задача может быть решена при помощи оптимизации, однако получаемая функция является дискретной и не обладает четко выраженными оптимумами. В связи с этим на первом этапе ограничим число степенных показателей тремя: -1, 0 и +1. В качестве статистических значений будем использовать семь показателей: $C_2(t_{i-1},t_i)$, $C_w(w_i,t_i)$, $C_a(a_{w_i},t_i)$, $C_1(t_i)$, $C_p(t_{i-1},p_{i-1},t_i)$, $C_{pa}(t_{i-1},p_{i-1},t_i,a_{w_i})$ и $C'_{pa}(t_{i-1},p_{i-1},t_i,a_{w_{i-1}})$. Таким образом, пространство поиска составило 3^7 = 2187 вариантов функции P.

Как это было показано ранее [Клышинский и др. 2015b], в текстах на русском языке порядка 40% слов могут быть не омонимичны. В связи с этим измерение точности снятия омонимии по всему тексту является практически значимой, но неточной мерой, так как показывает скорее степень однозначности слов в тексте, чем качество снятия омонимии. По этой причине мы замеряли два показателя – точность снятия омонимии на всём тексте (с тем, чтобы иметь возможность сравниваться с другими авторами) и только на омонимичных словах (с тем, чтобы оценить собственно качество снятия омонимии).

2.3 Результаты экспериментов

Для каждой функции было рассчитано две метрики: точность снятия неоднозначности на всём корпусе и только на словах, омонимичных по части речи и набору лексических параметров. Для экспериментов использовался синтаксически размеченных корпус СинТагРус [Apresjan, 2006; Богуславский, 2002], а также подкорпус НКРЯ [Lashevskaja, 2003] с морфологической разметкой, сделанной вручную.

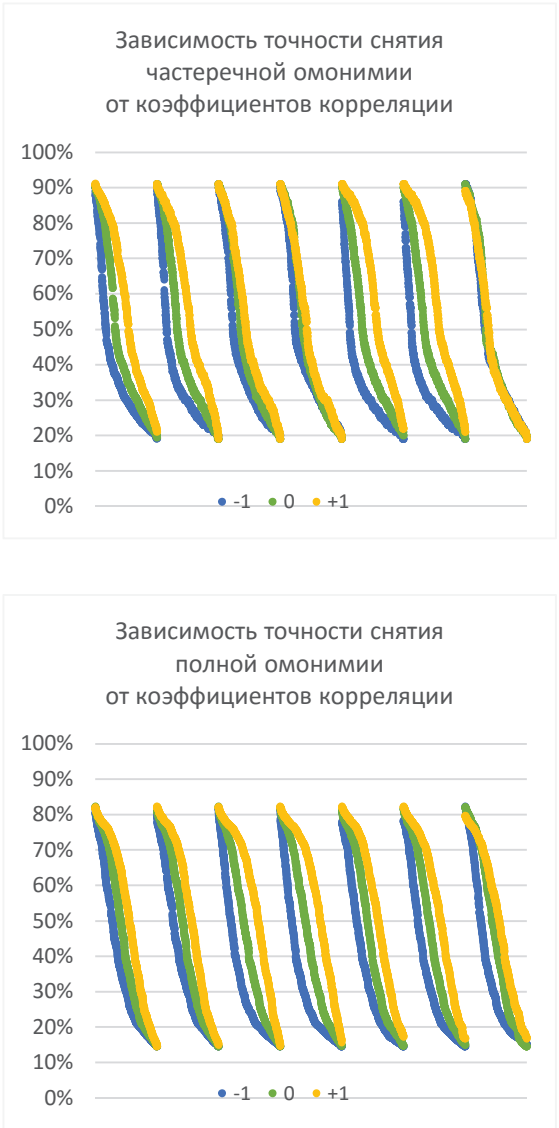

Рис. 1. Влияние коэффициентов корреляции на точность снятия омонимии. Слева направо расположены графики для счётчиков C_{pa}, C'_{pa}, C_p, C_2, C_w, C_a, и C_1.

На рис. 1 показаны результаты экспериментов на этих корпусах. Каждый из семи

графиков на рисунке показывает изменение качества омонимии при фиксированном показателе с различными степенями. На каждом графике представлены одни и те же данные, упорядоченные по убыванию.

Как видно из графиков, показатели поразному чувствительны к изменению степени в формуле (2).

Точность снятия частеречной омонимии всеми методами заняла диапазон от 19,0% до 91,2%, частеречная омонимия и омонимия по лексическим параметрам была снята с точностью от 14,5% до 82,2%.

При изучении коэффициентов корреляции лучших разметчиков оказалось, что наборы их коэффициентов существенно отличаются между собой, что ещё раз подтвердило нецелесообразность поиска коэффициентов методом оптимизации. Поскольку однозначный результат не было получен, было решено повторить эксперимент в другом пространстве параметров: были исключены значения, не использованные лучшими разметчиками, а для увеличения вариации были добавлены коэффициенты 0,5 и 2, что увеличило число вариантов до 37632.

Новые результаты расчётов выявили несколько разметчиков, показавших высокую точность снятия омонимии. Показатели лучших из них отражены в таблице 1.

Табл. 1. Показатели методов снятия омонимии

	Специализация метода	
	Частеречная омонимия	**Полная омонимия**
Значения	Степенные показатели	
$C_1(t_i)$	-1,0	-0,5
$C_2(t_{i-1}, t_i)$	0,5	1,0
$C_w(w_i, t_i)$	2,0	2,0
$C_a(a_{wi}, t_i)$	-1,0	1,0
$C_p(t_{i-1}, p_{i-1}, t_i)$	-0,5	0,5
$C_{pa}(t_{i-1},p_{i-1},t_i,a_{wi})$	1,0	2,0
$C'_{pa}(t_{i-1},p_{i-1},t_i,a_{wi-1})$	1,0	1,0
	Точность снятия омонимии	
Только омонимы	92,749%	88,040% *88,230%**
Весь корпус	**98,009%**	**93,544%** ***93,647%****

*с предварительной частеречной разметкой.

3 Влияние частотности слова на снятие омонимии

Как это было показано в наших предыдущих работах [Клышинский, 2015а], распределение слов по типам омонимии зависит от частотности этих слов. В связи с этим была высказана гипотеза, что учет данного параметра может повысить качество снятия омонимии. Косвенно подтверждением данной гипотезы может являться тот факт, что частота начальной формы $C_w(w_i, t_i)$ учитывается в полученной формуле во второй степени, то есть влияет на результат значительно больше других показателей.

Заметим, что если слово встречается слишком редко, статистический показатель будет возвращать сглаженное значение, что внесет шум в итоговый результат. В связи с этим мы также решили разделить статистику для высоко- и низкочастотных слов. Слова, частотность которых лежит ниже этого порога будут обрабатываться по формуле (2). Для наиболее частотных слов будет использоваться набор функций-счётчиков $\bar{C}$, принимающих дополнительный параметр: лемма.

Мы провели еще одну серию экспериментов, в которой изменялось количество слов, воспринимаемых как высокочастотные. Порог отсечения варьировался в пределах от 1 до 5000. Результаты оценки точности снятия омонимии показана на рисунке 2.

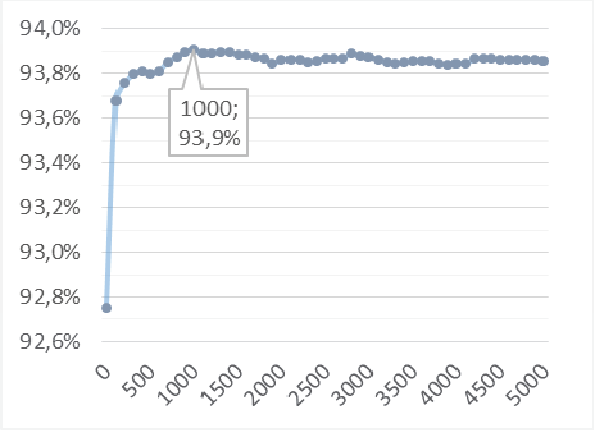

Рис. 2. Влияние порога частоты на точность снятия омонимии.

С учётом частотности, точность снятия частеречной омонимии составила 93,910% для омонимов и **98,327%** для всего корпуса.

4 Заключение

Представленные результаты работы методов снятия омонимии показывают, что пределы точности автоматической разметки ещё выше, чем можно было ожидать. Пространство коэффициентов корреляции исследовано далеко не полностью, и порог частоты также предполагает некоторую свободу выбора, так

что впереди ожидаются новые исследования и, скорее всего, новые результаты.

Благодарности

Данная работа выполнена при финансовой поддержке гранта РГНФ № 15-04-12019.

Список литературы

Apresjan J., Boguslavsky I., Iomdin B., et al. 2006 *A syntactically and semantically tagged corpus of Russian: State of the art and prospects* // In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006), May 24-26, 2006, Genoa, Italy.

Lashevskaja, O., Plungian V. 2003 *Morphological annotation in Russian National Corpus: a theoretical feedback* // Proc. Of 5th International Conference on Formal Description of Slavic Languages (FDSL-5). Nov. 2003, pp. 26-28.

Богуславский И.М., Иомдин Л.Л., и др. 2002 *Разработка синтаксически размеченного корпуса русского языка* // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» // СПб, изд-во Санкт-Петербургского университета, 2002, сс. 40–50.

Клышинский Э.С., Логачева В.К., Мансурова О.Ю. и др. 2015a *Исследование неоднозначности употребления слов в европейских языках* // Препринты ИПМ им. М.В. Келдыша. №4. 31 с.

Клышинский Э.С., Логачева В.К., Нивре Й. 2015b *Распределение неоднозначных слов в некоторых европейских языках* // Сборник трудов международной конференции «Корпусная лингвистика 2015». СПб. с. 270-277.

Рысаков С.В., Клышинский Э.С. *Статистические методы снятия омонимии* // Материалы восемнадцатого научно-практического семинара «Новые информационные технологии в автоматизированных системах». М., 2015.

Doug Cutting et al.. 1992. *A practical part-of-speech tagger*. Proceedings of the 3rd conference on Applied natural language processing, Stroudsburg, PA.