МОДЕЛИ ОБРАЗОВАНИЯ ЧИСЛИТЕЛЬНЫХ В РУССКОМ ЯЗЫКЕ

Юлия Кузнецова

Введение

Система InBase - это естественно-языковая оболочка для баз данных. Запросы к базам в InBase возможны на естественном языке. При анализе запросов возникают разнообразные проблемы, связанные с преобразованием запроса на формальный язык. Некоторые такие проблемы встречаются при анализе естественно-языковых запросов, в которых встречаются числительные. Для дальнейшей обработки необходимо преобразовать их в числовой формат.

Для обработки таких числительных необходимо подробно описать способы образования числительных и конструкций с числительными в языке.

Модели образования числительных

Модель образования числительных в русском языке можно описать как формулу

$[(P_{N,3}*100 + P_{N,2}*10 + P_{N,1}) *1000+(P_{N-1,3}*100 + P_{N-1,2}*10 + P_{N-1,1})]*1000^{N-2}+...$

$...+ (P_{2,3}*100 + P_{2,2}*10 + P_{2,1})*1000 +(P_{1,3}*100 + P_{1,2}*10 + P_{1,1})$

где $P_{N,M}$ - переменная, $0 \le P_{N,M} < 10$.

Например, число 252 921 674 532 можно представить как числовое выражение $[(2*100+5*10+2)*1000+(9*100+7*10+1)]*1000^2+(6*100+5*10+4)*1000+(5*100+3*10+2)$

Образование числительного в русском языке происходит аналогичным образом:

[(две*сти+пять*десят+две)*тысячи+(девять*сот+семь*десят+один)]*миллион+
(шесть*сот+семь*десят+четыре)*тысячи+(пять*сот+три*дцать+два)

При описании моделей образования числительных необходимо учитывать, что, в принципе, в русском языке возможно образование сколь угодно большого числительного. Образование числительного будет происходить по схеме образования числительных. Для описания модели образования числительных можно использовать ограниченное количество словоформ.

Необходимо заметить, что большая часть числительных обладает нестандартной морфологией, таким образом для обработки числительных необходимы записи о всех падежных формах числительного, а также информация о роде (для числа *два*: ***два*** *мальчика*, но ***две*** *девочки*)

Сложное числительное

Будем называть числительное, которое соответствует числу вида $P_{N,M}*L^{K,}$ сложным числительным, если согласно правилам русского правописания это число пишется в два или более слов, из которых одно соответствует L^K, а остальные соответствуют $P_{N,M}$. Таким образом *двести* не будет сложным числительным, а *две сотни* будет сложным числительным, для которого $P_{N,M}$ =2, L =10, а K=2. Назовем числа, которые могут выступать как $P_{N,M}$ - классом образующих, а L^K - классом порядковых чисел. Соответствующие им числительные будем называть соответственно образующими и порядковыми числительными.

К классу образующих может относиться любое числительное из таблицы словоформ, приведенной выше. К классу порядковых числительных традиционно относятся названия степеней 10 - *десяток*, *сотня*, *тысяча* и т.д. Но возможны и другие варианты, см. далее.

Специальные названия чисел

В качестве порядковых числительных от чисел 1000 и 1000 000 используются стандартные их наименования: *тысяча* и *миллион*, но для остальных степеней 10 в сложных числительных используются другие специальные названия: *десяток*, *сотня*.

Также к этому же классу числительных, необходимых для обработки, хотелось бы причислить слова *единица*, *двойка*, *тройка*, *четверка*, *пятерка*, *шестерка*, *семерка*, *восьмерка*, *девятка*, *десятка*, а также *дюжина*, *двадцатка* и *тридцатка*, о которых речь пойдет ниже.

Таблица 1.Таблица специальных названий чисел

	Стандарт	Специальное название
1	один	единица
2	два	двойка
3	три	тройка
4	четыре	четверка
5	пять	пятерка
6	шесть	шестерка
7	семь	семерка
8	восемь	восьмерка
9	девять	девятка
10	десять	десяток, десятка
12	двенадцать	дюжина
20	двадцать	двадцатка
30	тридцать	тридцатка
100	сто	сотня

Нестандартные представители классов образующих и порядковых числительных

Слова *пол* и *половина* несмотря на то, что у них нет соответствующих им числовых эквивалентов необходимо также внести в класс образующих числительных. Аналогично другим словам из этого множества слова *пол* и *половина* могут участвовать в образовании сложных числительных в качестве образующих числительных: *пол тысячи*, *полтора десятка*.

Слова *дюжина*, *двадцатка*, *тридцатка* примечательны тем, что могут употребляться как слова относящиеся к классу порядковых числительных: *две дюжины*, *две двадцатки*.

Эллипсис порядкового числительного при повторении

При повторении числительных возможна ситуация опущения порядкового числительного и упоминания только числительного класса образующих во втором употреблении: *сотня или 2, полторы-две сотни, от сотни до 3-ех.*

Для обработки таких конструкций можно использовать механизм образования таких чисел. Здесь важно, что конструкции *сотня-две* и *полторы-две сотни* никак нельзя понять как $100<X<2$ или $1,5<X<200$ соответственно. Первое вообще бессмысленно, второе математически осмысленно, но практически несравнимо.

Все такие конструкции, какими бы разными они внешне не казались (*дороже двух тысяч, но дешевле трех*; *больше трех сотен и меньше четырех*; *от пяти до шести миллионов*) всегда задают интервал на некоторой шкале. Таким

образом, после преобразования, которое приведет нашу числовую запись к виду A < X < B, необходимо произвести ряд проверок и если

1) одно из числительных, соответствующих A и B, является сложным числительным (назовем его S), а другое простым (назовем его T)

2) $S = P_1 * L^K$ и $T = P_2$, таким что P_1 и P_2 являются числами одного порядка (2 и 5, 20 и 50, 200 и 500)

3) при замене числа T на $T * L^K$ интервал для числа X не является вырожденным

(не $30 < X < 20$)[1], то в этом случае необходимо производить домножение простого числа на порядковое.

Таким образом, обработка фразы "пять-шесть сотен" будет происходить следующим образом. Сначала произойдет преобразование в числовой интервал: $5 < X < 600$. Затем, так как

1) числительное "пять" является простым, а числительное "шесть сотен" сложным

2) числа 5 и 6 являются числами одного порядка

3) интервал $500 < X < 600$ является осмысленным

в этом случае необходимо будет преобразование границ интервала: "пять-шесть сотен" = $500 < X < 600$.

Ошибочное написание

Также существует проблема достаточно частых ошибок в сложных числительных. Возможно, их следует учитывать при анализе.

Таблица 2. Таблица ошибочных написаний числительных

	Твор.П.	Ошибки в Твор.П.
50	пятьюдесятью	пятидесятью
60	шестьюдесятью	шестидесятью
70	семьюдесятью	семидесятью
80	восемьюдесятью/восьмьюдесятью	восьмидесятью
100	ста	стами
200	двумястами	двухстами
300	тремястами	трехстами
400	четырьмястами	четырехстами
500	пятьюстами	пятистами
600	шестьюстами	шестистами
700	семьюстами	семистами
800	восьмьюстами/восемьюстами	восьмистами
900	девятьюстами	девятистами
1000	тысячей/тыщей	тысячью

Анализ частых ошибок в написании числительных был произведен при помощи поисковой системы "Yandex".

Примечание о порядке сбора ошибочных употреблений

Анализ частотности того или иного ошибочного написания числительного производился на основе статистики системы "Яндекс", выданной на соответствующий запрос. Корпус текстов русского Интернета достаточно объемен, и хотя содержит ошибки, в значительной мере диагностичен даже в тех же ошибочных употреблениях. Например, слово «восемьдесят» имеет согласно словарю Зализняка в творительном падеже два варианта написания «восьмьюдесятью» и «восемьюдесятью». В следующей таблице приведены количества употрблений в русском Интернете этих вариантов, а также одного ошибочного, но используемого в разговорной речи варианта

[1] Ср. "*от трех до двух сотен*" = $3 < X < 200$, "*от двух до трех сотен*" = $200 < X < 300$

«восьмидесятью», варианта «восмьюдесятью», появляющегося при опечатке в одном из правильных вариантов, и совершенно невозможного ни в устной, ни в письменной речи варианта «восемидесятью».

Таблица 3. Статистика ошибочных употреблений

Словоформа	Количество употреблений
восемьюдесятью	83343
восьмьюдесятью	86259
восмьюдесятью	181
восьмидесятью	54
восемидесятью	0

Приведенная в таблице статистика кажется нам весьма показательной.

Литература

1. Барулин, А.Н. К построению модели синтеза русских нумеративов (глубинное и поверхностно-семантическое представление) // Московский лингвистический журнал т.2 М.,1996.
2. Зализняк А.А. Грамматический словарь русского языка: Словоизменение.-М.:Рус.яз.,1980.