Нормализация цифровой записи числительных с помощью аппарата условных случайных полей в русскоязычных текстах

Романенко Александр Александрович

Студент

Факультет ФУПМ МФТИ(ГУ), Москва, Россия

E-mail: angiff07@gmail.com

Одной из частых задач, с которыми сталкиваются на этапе синтеза речи, является правильное прочтение машиной так называемых «нестандартных» слов, то есть аббревиатур, сокращений, а также числительных, представленных в предложении цифрами [2,3]. В работе рассматривается задача определения грамматической формы числительных, записанных цифрами.

В отличие от других работ по данной теме [3,4], в качестве признаков в задаче машинного обучения предлагается использовать не только частотные характеристики данных, но и грамматические свойства контекста употребления числительного. Более того, в [4] рассматривалась задача нормализации только для количественных числительных. В этой же работе предлагается определить и тип числительного (TYPE — порядковое или количественное), а также падеж (CASE), род (GEN), число (SNGL) и одушевленность (ANIM).

В качестве алгоритма машинного обучения в работе используется модицированная линейная модель условных случайных полей (linear-chain CRF), изображенная на рис. 1. Заметим, что грамматические метки числительного представляются в виде цепочки целевых переменных: TYPE-CASE-GEN-SNGL-ANIM. Также отметим, что Begin, End — фиктивные вспомогательные переменные. Определив грамматические характеристики числительного, его цифровую запись можно однозначно перевести в словесную запись с помощью, например, конечного автомата.

В качестве множества прецедентов использовалось подмножество Национального корпуса русского языка [1] (10268 фраз, содержащих числительные). 8251 фраза использовалась для обучения, 2017 — для контроля. Наилучшая конфигурация признакового описания дает на тестовом множестве аккуратность $Acc = 92{,}39\%$. Для сравнения, в [4] аккуратность работы алгоритма, не использующего грамматические метки слов из контекста, составляет 86%. На рис. 2 приведены значения точности P, полноты R и F_1-меры, усредненные по

категориям.

Также на этих данных была проведена процедура кросс-валидации (5-fold CV). Оценка скользящего контроля $CV = 92{,}21\%$, что говорит о высокой обобщающей способности используемого метода.

Иллюстрации

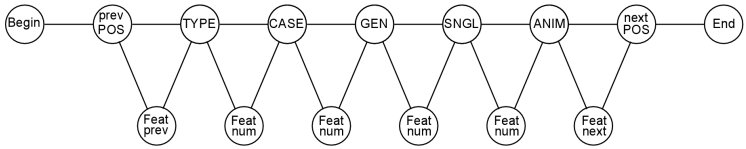

Рис. 1: Используемая модель CRF.

Мера качества	TYPE	CASE	GEN	SNGL	ANIM
P	97,21	91,33	89,77	82,39	87,66
R	97,21	92,93	90,74	85,97	95,05
F_1	97,21	92,10	90,24	84,05	91,11

Рис. 2: Результаты работы алгоритма, усредненные по категориям.

Литература

1. Сайт Национального корпуса русского языка:
   `http://www.ruscorpora.ru/`
2. Olinsky C., Black A. W. Non-standard word and homograph resolution for asian language text analysis // In proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), ISCA, 2000, P. 733–736.
3. Sproat R., Black A. W., Chen S. F., Kumar Sh., Ostendorf M., Richards C. et al. Normalization of Non-Standard Words // Computer Speech & Language, Vol. 15, 2001, P. 287–333.
4. Sproat R. Lightly supervised learning of text normalization: Russian number names // Workshop on Language Spoken Technology (SLT), IEEE, 2010, P. 436–441.