ОДИН ПОДХОД К ПРЕДСКАЗАНИЮ МОРФОЛОГИЧЕСКИХ ХАРАКТЕРИСТИК ОТСУТСТВУЮЩИХ В СЛОВАРЕ СЛОВ

Ф.А. Николаев, В.Д. Соловьев
Казанский федеральный университет
Казань

При морфологическом анализе текстов проблемой является обработка слов, отсутствующих в словарях. В ряде категорий текстов – научные статьи, страницы интернет-форумов — число таких слов достигает 20%. Это, прежде всего, имена собственные и разнообразные аббревиатуры, а также редкие и авторские (особенно, в произведениях футуристов и фантастов) слова. Некоторая классификация таких слов приведена в [1].

В данном исследовании реализована следующая идея. Основная морфологическая информация (число, род, падеж и др.) выражается в конце слова и, соответственно, анализируя конец неизвестного (отсутствующего в словаре) слова можно с высокой точностью предсказать его морфологические характеристики. В соответствие с принятой в информатике терминологией, конец слова (любой длины) будем называть суффиксом. Наш алгоритм суффиксы неизвестного слова сопоставляет с суффиксами слов из словаря OpenCorpora (построенного на базе известного словаря Зализняка) и находит наиболее длинный совпадающий суффикс. Важная идея состоит в том, что чем длиннее совпадающий суффикс у двух слов, тем они имеют больше общих морфологических характеристик. Для поиска слов в словаре использована структура данных в виде троичного дерева. На рис. 1 приведен простой алгоритм на Java поиска слов с наиболее длинным совпадающим суффиксом.

```
    public List<Wordform> getSimilarEntries(String str) {
     // Цикл по началу суффикса (от второй буквы до последней)
     for (int beginIndex = 1; beginIndex < str.length() - 1; beginIndex++){
           String suffix = str.substring(beginIndex);
           // Извлечение словоформ с требуемым суффиксом
           List<Wordform> similars = ImmutableList.copyOf(
                 wfByString.matchSuffix(suffix)
           );
           if (similars != null && !similars.isEmpty()) {
                 // Возвращаем найденные для самого длинного
                 // суффикса словоформы
                 return similars;
           }
     }
     return null;
}
```

Рис. 1. Алгоритм поиска словоформ с длиннейшим совпадающим суффиксом

В финальной реализации он был оптимизирован за счет элиминации циклического просмотра всех суффиксов и объединения процедур поиска суффикса и предсказания.

Важное отличие с другими работами по данной тематике, например, с [2, 3] заключается в том, что при реализации используется единый словарь, тогда как в [2] [3] выделяются отдельно словарь основ и словарь окончаний. К сожалению, в работе [3] отсутствуют оценки эффективности, что не позволяет сравнить подходы количественно.

Эффективность алгоритма оценивалась с помощью принятой в информационном поиске оценки аккуратности [4]. Тестирование предложенного алгоритма осуществлено двумя способами.

Во-первых, взята дорожка «Редкие слова» с форума «Оценка методов автоматического анализа текстов» [1]. Аккуратность оказалась равна 0,68, что близко к результатам участников данной дорожки.

К сожалению, для этой дорожки было подготовлено всего 75 слов. Для получения большей статистики мы обратились к словарю OpenCorpora.

Второй подход к тестированию состоял в том, что из словаря OpenCorpora выделено порядка тысячи лемм каждой из основных частей речи (существительные, полные и краткие прилагательные, компаративы, глаголы, инфинитивы, полные и краткие причастия, деепричастия, наречия), были выделены их

словоформы, которых в итоге получилось 95364, и их характеристики определены на основе остальных слов словаря.

Также применялся новый подход к оценке, являющийся более гибким и учитывающий при оценке правильность предсказания отдельных граммем. При этом анализ считался правильным только при полном совпадении всех найденных системой морфологических характеристик с указанными в словаре. Не найденные и лишние характеристики приводили к штрафу при оценке. Ошибка в определении части речи штрафовалась с повышенным весом.

Более точно этот подход можно выразить формулой:

Accuracy = t_p / (t_p + f_n + f_p), где

t_p — количество граммем, предсказанных правильно,

f_n — количество лишних граммем (нет в словаре, есть на выходе системы),

f_p — количество недостающих граммем (есть в словаре, нет на выходе системы).

Аккуратность представленного морфоанализатора при подсчете данным способом составила 0,52. Отдельный интерес может представлять также статистика аккуратности, собранная по отдельным частям речи. Для существительных она составляет 0,57, для глаголов — 0,68, для инфинитивов — 0,75, для полных и кратких имен прилагательных — 0,47 и 0,36 соответственно.

Работа выполнена по государственному заданию высшим учебным заведениям от Министерства образования и науки РФ (№ 8.3358.2011).

ЛИТЕРАТУРА

1. Оценка методов автоматического анализа текста: морфологические парсеры русского языка / О. Ляшевская, И. Астафьева, А. Бонч-Осмоловская, А. Гарейшина, Ю. Гришина, В. Дьячков, М. Ионов, А. Королева, М. Кудринский, А. Литягина, Е. Лучина, Е. Сидорова, С. Толдова, С. Савчук, С. Коваль // Компьютерная лингвистика и интеллектуальные технологии. Вып. 9 (16). М.: РГГУ, 2010. С. 318—326. URL: http://www.dialog-21.ru/digests/dialog2010/materials/pdf/49.pdf (дата обращения: 15.08.2013).
2. Сокирко А. В. Морфологические модули на сайте www.aot.ru // Диалог’2004: тр. Междунар. конф. М.: Наука, 2004. С. 559—564. URL: http://aot.ru/docs/sokirko/Dialog2004.htm (дата обращения: 15.08.2013).
3. Кузнецов И. П., Сомин Н. В. Особенности лексико-морфологического анализа при извлечении информационных объектов и связей из текстов естественного языка // Компьютерная лингвистика и интеллектуальные технологии. Вып. 9 (16). М.: РГГУ, 2010. С. 254—264. URL: http://www.dialog-21.ru/digests/dialog2010/materials/pdf/40.pdf (дата обращения: 15.08.2013).
4. Accuracy and precision [Электронный ресурс] // Wikipedia [сайт]. URL: https://en.wikipedia.org/wiki/Accuracy_and_precision (дата обращения: 15.08.2013).