В.Л. Бобичев

АВТОМАТИЧЕСКОЕ СНЯТИЕ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ ПРИ РАЗМЕТКЕ КОРПУСА

1. Введение

Автоматический морфологический анализ текста является одним из основных этапов предварительной обработки при решении большинства задач автоматической обработки текстов. На базе этого этапа осуществляется синтаксический и/или семантический анализ.

Задача морфологической разметки слов текста не является тривиальной по причине морфологической неоднозначности слов, а также наличия в тексте слов, которые отсутствуют в словаре.

Проблема снятия морфологической многозначности решалась разными способами. Первые созданные алгоритмы были основаны на правилах[1]. Позже для решения этой задачи были применены статистические алгоритмы. Уже классическим считается метод, основанный на модели Маркова. Один из методов такого типа был применен для морфологической разметки румынского текста[2].

2. Алгоритм PPM

Статистические методы, применяемые для решения задач автоматической обработки текстов, опираются на статистические закономерности текста. Статистическая модель PPM (Prediction

---

[1] *Tufiş D., Popescu O.* A Knowledge-Based Approach to Morpho-lexical Processing of Natural Language, in Proceedings of the International Conference for Young Computer Scientists, Beijing, 1991.

[2] *Tufiş D., Mason O.* Tagging Romanian texts: A Case Study for QTAG, a Language Independent probabilistic tagger, First International Conference on Language Resources and Evaluation, Granada, 28–30 May, 1998.

by Partial Matching – предсказание по частичному совпадению) была создана для сжатия текстов и считается оптимальной методикой сжатия текстов[1]. PPM использует вариант смешивания условных вероятностей в зависимости от контекстов разной длины, для оптимального определения вероятности.

Логично было предположить, что созданная PPM модель может быть использована для решения задач автоматической обработки текста на естественном языке[2]. Модель PPM может быть создана как на базе букв текста, так и на базе слов, и других последовательных элементов текста. В работе была применена модель PPM на базе морфологических кодов.

3. Морфологическая разметка

Как было отмечено выше, морфологическая разметка текста обычно осуществляется в два этапа. Результатом первого этапа разметки с помощью морфологического словаря является текст, слова которого сопровождаются подробным описанием, состоящим из двух частей. Первая часть – это исходная форма слова (лемма), а вторая – так называемый морфологический код, краткая аббревиатура, описывающая морфологические характеристики слова. Для разметки были использованы морфологические коды, разработанные проектом MULTEXT-EAST[3]. Слова с морфологической многозначностью помечены всеми возможными кодами, найденными в словаре. Кроме того, некоторые слова текста не были найдены в словаре и были помечены в тексте как неизвестные (unknown).

---

[1] *Moffat A.* Implementing the PPM data compression scheme. IEEE Transactions on Communications. 1990. Vol. 38. № 11. P. 1917–1921.

[2] *Teahan W.J.* Modelling English Text. PhD thesis, University of Waikato. 1998. P. 243.

[3] MULTEXT-East lexical specifications. Concede Edition. Available at **http://nl.ijs.si/ME/V2/msd/** Site visited 20.04.2008.

Задачей следующего этапа работы является снятие морфологической многозначности, а именно, выбор одного, верного кода для каждого слова. Для многозначных слов чаще всего представлено два или три кода, из которых необходимо выбрать один. Наибольшую проблему представляют неизвестные слова. Для того чтобы выбрать верный код для них приходится перебирать весь набор кодов. В нашем алгоритме мы намеренно уменьшили набор кодов, перебираемых в случае неизвестных слов с целью ускорения работы программы.

Примененный метод оценивает вероятности кодов на базе корпуса, проверенного вручную, используя метод PPM, затем автоматически снимает многозначность слов в тексте, используя модифицированный вариант алгоритма Витерби.

4. Результаты экспериментов

Используя метод PPM с максимальным контекстом равным четырем предыдущим кодам, были определены вероятности кодов на базе части корпуса откорректированного вручную. Затем было осуществлено снятие морфологической многозначности в текстах, размеченных автоматически. Анализ полученных результатов в сравнении с ручной разметкой представлен в табл. 2.

Анализируя результаты, представленные в табл. 2, можем сказать, что процент ошибок равный в среднем 4,7% не столь уж и мал. Например, если процент правильно помеченных слов равен 95%, это значит, что каждое двадцатое слово помечено неверно. Таким образом, если средняя длина предложения в тексте составляет 20 слов, каждое предложение текста будет содержать примерно одно неверно помеченное слово, которое в свою очередь повлечет за собой ошибки при синтаксическом анализе. Однако во многих случаях часть речи слова была определена верно, как и другие его характеристики, и ошибка была допущена лишь для какой-либо одной из характеристик слова. К примеру,

для существительного неверно был определен род или число, а все остальные характеристики были определены верно. Поэтому в табл. 2 была добавлена последняя колонка, которая отражает процент ошибок в частях речи. Видно, что процент этих ошибок практически в два раза меньше – 2,5%. Таким образом, процент правильно определенных частей речи равен 97,5%. Слова с правильно определенной частью речи не спровоцируют ошибок при последующем синтаксическом анализе.

Таблица 2. Процент верно определенных кодов и ошибок после автоматического снятия многозначности

№ текста	**Процент правильных кодов**	**Всего ошибок**	**Ошибок в определении части речи**
10	96,7%	17 (3,3%)	7 (1,4%)
63	96,6%	5 (3,4%)	5 (3,4%)
151	93%	42 (7%)	25 (4,2%)
205	88,8%	56 (11,2%)	17 (3,4%)
503	97,8%	13 (2,2%)	3 (0,5%)
576	94,9%	15 (5,1%)	10 (3,4%)
599	95,5%	31 (4,5%)	22 (3,2%)
630	97%	18 (3%)	13 (2,2%)
655	95,1%	12 (4,9%)	6 (2,5%)
777	97,7%	4 (2,3%)	2 (1,2%)
среднее	95,3%	4,7%	2,5%

Последующий более детальный анализ ошибок показал, что в некоторых случаях для многозначных слов верный вариант кода просто не был представлен в исходном списке кодов. Наиболее частый случай – разметка причастий, которые, встречаясь с существительным, считаются прилагательными: «*construit*» – *построенный*. Часть таких слов в словаре встречается с двумя кодами, и система может выбрать правильный вариант, часть же встречается только с кодом причастий. При ручной корректи-

ровке разметки вместо кода причастия был введен код прилагательного. Система же вообще не анализирует другие варианты кодов, если слово изначально было помечено только одним кодом. Такого рода ошибки составляют больше трети из общего числа ошибок, допущенных системой.

Следует отметить, что процент ошибок при ручной корректировке почти в пять раз меньше, чем при автоматической. Это объясняется тем, что люди при необходимости исправляли первичную разметку и даже текст, если по смыслу было видно что слово написано с ошибкой (опечатки, ошибки сканирования). Очевидно, что система снятия морфологической многозначности не в состоянии производить такие действия.

5. Выводы

В данной работе была представлена методика снятия морфологической многозначности слов в тексте с использованием модели PPM (Prediction by Partial Matching – предсказание по частичному совпадению). Полученные результаты показывают, что данная методика может быть использована для разметки корпуса. Метод верно определяет часть речи для 97,5% слов в тексте, что сопоставимо с другими методами, применяемыми для данной задачи.

Необходимо отметить, что качество снятия морфологической многозначности слов зависит и от качества текстов, и от полноты словаря и от набора морфологических характеристик, которыми помечаются слова.

Полученные результаты могут быть улучшены за счет качества первоначальной разметки, а именно, улучшения словаря. Кроме того, возможно встроить модуль угадывания морфологических характеристик неизвестных слов на основе их окончаний.