УДК [004.934:519.21/.24]:[81’322:811.161.1’36]
ББК [32.973:22.17]:[81.1:81.411.2-21]

А. А. Порохнин

АНАЛИЗ СТАТИСТИЧЕСКИХ МЕТОДОВ СНЯТИЯ ОМОНИМИИ В ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

A. A. Porokhnin

ANALYSIS OF STATISTICAL METHODS OF A PART-OF-SPEECH DISAMBIGUATION IN RUSSIAN TEXTS

Омонимия осложняет автоматическую обработку текста. Для текстов на английском языке достаточно широко представлены методы снятия омонимии, основанные на использовании вероятностной модели, которые дают достаточно высокую точность. Проблема для текстов на русском языке заключается не только в частеречной омонимии, свойственной текстам на английском языке, но и в морфологической и лексической омонимии. Ввиду того, что составление математической модели для русского языка, который отличается свободным расположением слов в предложении, затруднено, для снятия омонимии в текстах на русском языке большее развитие получили методы, основанные на правилах. В целях выявления результатов работы метода опорных векторов и скрытой марковской модели для снятия частеречной и полной омонимии при обработке текстов на русском языке, проводится эксперимент, в ходе которого используется подкорпус со снятой омонимией национального корпуса русского языка. Показано, что скрытая марковская модель для снятия омонимии в текстах на русском языке работает лучше метода опорных векторов.

Ключевые слова: омонимия, частеречная омонимия, морфологическая омонимия, лексическая омонимия, методы снятия омонимии, скрытая марковская модель, метод опорных векторов.

Ambiguity complicates text processing. As for English texts there are a number of disambiguation methods based on application of the probability method, which gives high precision results. Regarding to the Russian texts the problem is not only in part-of-speech ambiguity specific for English texts, but also in morphological and lexical ambiguity. In view of the fact that it is difficult to create a mathematical model for Russian language with free order of words in a sentence, the disambiguation methods based on the rules have received a larger development. In order to define the results of support vectors method and hidden Markov’s model of a part-of-speech disambiguation and full disambiguation in Russian texts processing the experiment, which lies in the use of sub-corpus with the disambiguated national corpus of the Russian language, is set up. It is shown that the hidden Markov’s model for disambiguation in Russian texts works better than the method of support vectors.

Key words: ambiguity, part-of-speech ambiguity, morphological ambiguity, lexical ambiguity, disambiguation methods, hidden Markov’s model, method of support vectors.

Введение

С момента изобретения компьютера человек стремится к тому, чтобы компьютер мог понимать человека. Разрабатываются все новые системы для распознавания текста, речи, жестов. Автоматическая обработка текста – одно из наиболее старых направлений исследований в этой области. Первые работы по автоматической обработке текстов относятся к 50-м гг. XX столетия.

Автоматическая обработка текста делится на несколько этапов, одним из которых является морфологический. На данном этапе для каждого слова определяются морфологические характеристики (род, число, падеж, склонение, залог и др.) и начальная форма слова, называемая леммой [1]. Задача морфологической разметки осложняется омонимией. Для английского языка, который является флективнобедным, существует только частеречная омонимия. Однако даже такой вид омонимии представляет большую проблему при автоматической обработке текста. Эта проблема кроется в словах [2], которые по своей гибридной природе имеют свойства нескольких частей речи, что затрудняет их отнесение к той или иной группе слов. Таким образом, разработка совершенной классификации частей речи остается вечной и вряд ли разрешимой проблемой.

Для русского языка, как для флективнобогатого, с развитой морфологией, проблема омонимии значительно сложнее, т. к. в русском языке можно выделить несколько видов омонимии:

1. Частеречная омонимия – омонимия, при которой одно и то же слово относится к различным частям речи:

Мальчик, моя$^{\text{дееп.}}$ полы, разлил ведро.

Моя$^{\text{мест.}}$ сумка стояла на кровати.

В данном примере слово «моя» является деепричастием в первом предложении и местоимением – во втором.

2. Морфологическая омонимия – омонимия, при которой одно и то же слово является одной частью речи, но относится к разным леммам и совпадает лишь в некоторых формах:

Данная история отложилась в веках$^{\text{век}}$.

Снег медленно таял на её веках$^{\text{веко}}$.

В данном примере слово «веках» относится к разным леммам. В первом предложении для слова «веках» леммой является слово «век». Во втором предложении – слово «веко».

3. Лексическая омонимия – омонимия, при которой слова, относящиеся к одной и той же части речи и имеющие одинаковые леммы, различаются лишь лексическим смыслом:

Вражеская армия напала на за́мок.
Он вставил ключ в замо́к.

В обоих предложениях слово «замок» имеет одинаковые грамматические характеристики, но разный смысл. В первом предложении слово «замок» означает строительное сооружение, во втором – охранное устройство.

В связи с вышеизложенным были поставлены следующие цели исследований:

– проанализировать методы, применяющиеся для снятия омонимии в текстах на английском языке;

– оценить возможности метода опорных векторов и скрытой марковской модели для снятия омонимии в текстах на русском языке.

Обзор методов снятия омонимии в текстах на английском языке

Существует большое число работ по снятию омонимии в текстах на английском языке, использующих вероятностные методы с довольно высокой точностью. В английском языке, в котором существует точный порядок следования слов в предложении, применение данных методов упрощено. Для наглядности приведем результаты работы этих методов по снятию омонимии в табл. 1.

Методы по снятию омонимии принято подразделять на статистические; методы, основанные на правилах и гибридные подходы.

Системы, построенные на основе статистических методов, снимают омонимию в текстах на этапе морфологического анализа, используя статистику совместной встречаемости грамматических признаков слов из корпусов, размеченных вручную.

Системы, работающие на основе правил, снимают омонимию на этапе синтаксического анализа. Разработка таких систем – довольно долгий и трудоемкий процесс, требующий привлечения экспертов в данной области.

Гибридные подходы построены на основе статистических методов и оснащены небольшим числом правил.

Таблица 1

Результаты работы методов снятия омонимии в текстах на английском языке

Метод	Точность определения частеречной принадлежности , %	
	всех слов	незнакомых слов
Скрытая марковская модель [3]	96,46	85,86
Метод опорных векторов [4]	97,16	89,01
Сеть циклической зависимости максимальной энтропии [5]	97,05	–
Двунаправленный простейший первый вывод максимальной энтропии [6]	97,15	–
Двунаправленное обучение перцептрона [7]	97,33	–
Перцептрон с заглядыванием вперед [8]	97,22	–

Возможно из-за большой сложности русского языка, со свободным порядком слов в предложении, возникло предположение, что невозможно описать русский язык математической моделью, и поэтому развитие систем для снятия омонимии в текстах на русском языке пошло по пути систем, основанных на правилах.

Метод опорных векторов и скрытая марковская модель

Для сравнения были выбраны два метода снятия омонимии в текстах на английском языке: метод опорных векторов и скрытая марковская модель (СММ). Достоинствами этих методов является возможность снятия частеречной омонимии для слов, которых нет в словаре.

Скрытая марковская модель. Скрытая марковская модель – статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами. Задачей ставится разгадывание неизвестных параметров на основе наблюдаемых переменных [9]. На рис. 1 графически представлена общая структура СММ.

Овалы представляют собой переменные со случайным значением. Случайная переменная $x(t)$ представляет собой значение скрытой переменной в момент времени t. Случайная переменная $y(t)$ – это значение наблюдаемой переменной в момент времени t. Стрелки на диаграмме символизируют условные зависимости.

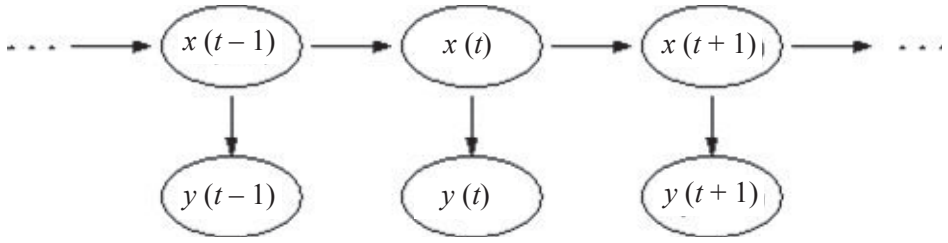

Рис. 1. Общая структура СММ

Становится ясным, что значение скрытой переменной $x(t)$ зависит только от значения скрытой переменной $x(t-1)$. Эта зависимость получила название «свойство Маркова». В то же время значение наблюдаемой переменной $y(t)$ зависит только от значения скрытой переменной $x(t)$.

В задаче снятия морфологической омонимии наблюдаемыми величинами являются слова, а скрытыми – морфологические тэги. Для возможности анализа слов, отсутствующих в словаре, используются трехбуквенные постфиксы, а решение задачи по определению морфологических признаков сводится к вычислению последовательности наиболее вероятных значений скрытых величин:

$$y'(t) = \operatorname{argmax} P(y|x).$$

Метод опорных векторов. Метод опорных векторов, предложенный в [10], является одной из наиболее популярных методологий обучения по прецедентам.

Основная идея метода опорных векторов – нахождение разделяющей поверхности с максимальным зазором между векторами различных классов. На рис. 2 графически представлен пример разделяющих плоскостей.

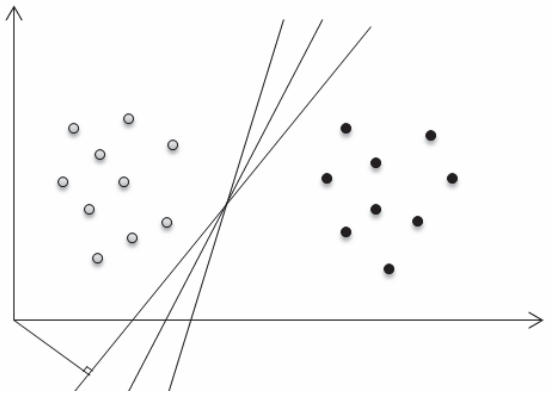

Рис. 2. Пример разделяющих плоскостей

В случае двух классов разделяющей поверхностью является гиперплоскость, которая делит пространство признаков на два полупространства. В случае большего числа классов разделяющая поверхность является кусочно-линейной.

Любая плоскость из n-мерного пространства R^n может быть представлена в виде

$$w\cdot\ x + b = 0,$$

где w – вектор ортогональной плоскости; b – коэффициент смещения плоскости.

Необходимо расположить разделяющую плоскость так, чтобы она максимально далеко отстояла от ближайших к ней точек классов, т. е. найти такие вектор w и число b, чтобы выполнялось следующее условие:

$$y = \text{sign}(w \cdot x + b).$$

Метод изначально разрабатывался для нахождения разделяющей плоскости между двумя классами объектов. В случае снятия частеречной омонимии имеют место более двух классов объектов, а именно количество классов, которое соответствует количеству частей речи.

Для снятия омонимии в текстах на русском языке используется такое число классов, которое соответствует тэгам омонимичных слов. Тэгом принято называть сокращение от грамматических характеристик слова.

В случае нескольких классов на практике часто применяется решающее правило, основанное на разбиении задачи на бинарные по схеме «один-против-остальных» (One-vs-Rest):

$$y = \text{argmax}\ (\langle w, x\rangle + b).$$

Для тренировки модели была использована библиотека LIBSVM, работающая с большой выборкой и с параметрами обучения «один-против-остальных».

Эксперимент

Для проведения эксперимента использовался подкорпус со снятой омонимией национального корпуса русского языка (НКРЯ), который разбивался на две части. На одной части подкорпуса производилось обучение модели, другая часть подкорпуса использовалась для тестирования работы методов снятия омонимии.

В ходе работы проводилось два эксперимента. В обоих случаях для обучения модели, как для метода опорных векторов, так и для СММ, использовался локальный контекст омонима. Под локальным контекстом в данной работе принимается не более двух соседних слева и справа тэгов слов.

Первый эксперимент проводился для снятия частеречной омонимии. Напомним, что именно этот вид омонимии присущ текстам на английском языке. Эксперимент проводился для выяснения точности, с которой данные методы могут снимать омонимию в текстах на русском языке.

Второй эксперимент проводился для определения дополнительного набора грамматических характеристик. К части речи добавлялись следующие грамматические характеристики: падеж, одушевлённость, число, лицо, время, наклонение. Остальные наборы грамматических характеристик, приписанные словам в НКРЯ, не учитывались.

Для различения лексических омонимов было введено понятие смысла. Под смыслом будем понимать строковую константу, точно описывающую предмет. Лексическая омонимия была определена только для существительных, которые можно различить по ударению. Например:

За'мок – строение

Замо'к – устройство

Оценка точности работы алгоритма

Методы снятия омонимии тестировались на части размеченного корпуса, на котором не проводилось обучение. Для омонима выбирались возможные тэги и лемма по словарю. Затем каждым методом определялся один наиболее вероятный тэг и сравнивался с тэгом из корпуса.

Для первого эксперимента омонимия считалась снятой, если верно была определена часть речи. Для второго эксперимента результат снятия омонимии засчитывался как правильный в том случае, если совпадали грамматические характеристики, описанные выше, и ударение как показатель снятой лексической омонимии.

Полученные результаты

Результаты, полученные в ходе эксперимента, приведены в табл. 2 и 3: в табл. 2 – результаты снятия частеречной омонимии, в табл. 3 – результаты снятия полной омонимии.

Таблица 2

Результаты работы методов снятия частеречной омонимии в текстах на русском языке

Метод	Точность определения частеречной принадлежности, %		
	всех слов	всех слов с учётом лексической омонимии	незнакомых слов
СММ	92,78	90,56	87,32
Метод опорных векторов	93,56	90,98	89,13

Таблица 3

Результаты работы методов снятия полной омонимии в текстах на русском языке

Метод	Точность определения тэгов, %		
	всех слов	всех слов с учётом лексической омонимии,	незнакомых слов
СММ	87,73	84,39	64,25
Метод опорных векторов	86,12	82,89	62,41

Анализ результатов и выводы

Таким образом, оба метода неплохо справляются со снятием частеречной омонимии в текстах на русском языке. Однако результаты снятия частеречной омонимии в текстах на русском языке ниже, чем в текстах на английском языке. С задачей снятия полной омонимии данные методы справляются гораздо хуже, чем с задачей снятия частеречной омонимии. Скрытая марковская модель для снятия омонимии в текстах на русском языке работает лучше метода опорных векторов.

В большей степени методы ошибаются при разметке местоимений, сокращений, инициалов и имен собственных. Кроме того, методы часто ошибаются с выбором падежных форм.

Известны и более высокие результаты для СММ. Так, в [11] исследователи группы автоматической обработки текста (АОТ), использовав специальную формулу сглаживания, сходную

с формулой из [12], смогли добиться в своем эксперименте точности в 97 % в снятии морфологической омонимии.

В [13] был описан метод, разработанный сотрудниками Яндекса, который по показателям снятия морфологической омонимии превзошел метод, разработанный исследователями группы АОТ и использующий СММ.

В дальнейшем планируется изучение и других методов снятия омонимии в текстах на английском языке, анализ работ по снятию омонимии в языках, близких по строению к русскому. При проведении эксперимента с новыми методами снятия омонимии планируется применять не только тэги, но и конкретные словосочетания и использовать пошаговое снятие омонимии соотносительно их типам. На первом этапе работы системы будет сниматься частеречная омонимия, на следующих этапах – морфологическая омонимия и затем лексическая.

СПИСОК ЛИТЕРАТУРЫ

1. *Большакова Е. И.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова: учеб. пособие. – М.: МИЭМ, 2011. – 272 с.

2. *Резунова М. В.* К проблеме частеречной классификации слов в языках / М. В. Резунова // Изв. Рос. гос. пед. ун-та им. А. И. Герцена. – 2005. – № 11. – С. 59–64.

3. *Brants Thorsten*. TnT – A Statistical Part-of-Speech Tagger / Brants Thorsten // 6th Applied Natural Language Processing Conference. – 2000.

4. *Giménez J.* SVMTool: A general POS tagger generator based on Support Vector Machines. // J. Giménez, L. Márquez // Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). – Lisbon, Portugal.

5. *Tsuruoka Yoshimasa.* Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics / Tsuruoka Yoshimasa, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, Jun'ichi Tsujii // 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382–392.

6. *Tsuruoka Yoshimasa.* Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data / Tsuruoka Yoshimasa, Jun'ichi Tsujii // Proceedings of HLT/EMNLP, 2005, pp. 467–474.

7. *Shen L.* Guided learning for bidirectional sequence classification / L. Shen, G. Satta, A. Joshi // Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), pp. 760–767.

8. *Tsuruoka Yoshimasa.* Learning with Lookahead: Can History-Based Models Rival Globally Optimized Models? / Tsuruoka Yoshimasa, Yusuke Miyao, Jun'ichi Kazama // Proceedings of the Fifteenth Conference on Computational Natural Language Learning, 2011, pp. 238–246.

9. *Ветров Д. П.* Скрытые марковские модели / Д. П. Ветров: http://www.machinelearning.ru /wikiimages/8/83/GM12_3.pdf.

10. *Вапник В. Н.* Теория распознавания образов / В. Н. Вапник, А. Я. Червоненкис. – М.: Наука, 1974. – 416 с.

11. *Сокирко А. В.* Сравнение двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) / А. В. Сокирко, С. Ю. Толдова // Интернет-математика-2005. Автоматическая обработка веб-данных. – М., 2005. – С. 80–94.

12. *Thede S. M.* A Second-Order Hidden Markov Model for Part-of-Speech Tagging / S. M. Thede, M. P. Harper // Proceedings of the 37th Annual Meeting of the ACL, 1999.

13. *Зеленков Ю. Г.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиции соседних слов / Ю. Г. Зеленков, И. В. Сегайлович, В. А. Титов // Компьютерная лингвистика и интеллектуальные технологии: тр. Междунар. семинара «Диалог-2005». – М.: Наука, 2005.

REFERENCES

1. Bol'shakova E. I., Klyshinskii E. S., Lande D. V., Noskov A. A., Peskova O. V., Iagunova E. V. *Avtomaticheskaia obrabotka tekstov na estestvennom iazyke i komp'iuternaia lingvistika* [Automated processing of texts in natural language and computer linguistics]. Moscow, MIEM, 2011. 272 p .

2. Rezunova M. V. K probleme chasterechnoi klassifikatsii slov v iazykakh [To the problem of part-of-speech classification of words in languages]. *Izvestiia Rossiiskogo gosudarstvennogo pedagogicheskogo universiteta im. A. I. Gertsena*, 2005, no. 11, pp. 59–64.

3. Brants Thorsten. TnT – A Statistical Part-of-Speech Tagger. *6th Applied Natural Language Processing Conference*, 2000.

4. Giménez J., Márquez L. SVMTool: A general POS tagger generator based on Support Vector Machines. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal, 2004.

5. Tsuruoka Yoshimasa, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, Jun'ichi Tsujii. Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics. *10th Panhellenic Conference on Informatics, LNCS 3746*, 2005, pp. 382–392.

6. Tsuruoka Yoshimasa, Jun'ichi Tsujii. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. *Proceedings of HLT/EMNLP*, 2005, pp. 467–474.

7. Shen L., Satta G., Joshi A. Guided learning for bidirectional sequence classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007),* 2007, pp. 760–767.

8. Tsuruoka Yoshimasa, Yusuke Miyao, Jun'ichi Kazama. Learning with Lookahead: Can History-Based Models Rival Globally Optimized Models? *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 238–246.

9. Vetrov D. P. *Skrytye markovskie modeli* [Hidden Markov's models]: http://www.machinelearning.ru/wiki/images/8/83/GM12_3.pdf.

10. Vapnik V. N., Chervonenkis A. Ia. *Teoriia raspoznavaniia obrazov* [Theory of image recognition]. Moscow, Nauka Publ., 1974. 416 p.

11. Sokirko A. V., Toldova S. Iu. Sravnenie dvukh metodik sniatiia leksicheskoi i morfologicheskoi neodnoznachnosti dlia russkogo iazyka (skrytaia model' Markova i sintaksicheskii analizator imennykh grupp) [Comparison of two methods of lexical and morphological disambiguation for Russian language (hidden Markov's model and synthetic analyzator of noun phrases)]. *Internet-matematika-2005*. *Avtomaticheskaia obrabotka veb-dannykh*. Moscow, 2005, pp. 80–94.

12. Thede S. M., Harper. M. P. A Second-Order Hidden Markov Model for Part-of-Speech Tagging. *In Proceedings of the 37th Annual Meeting of the ACL*, 1999.

13. Zelenkov Iu. G., Segailovich I. V., Titov V. A. Veroiatnostnaia model' sniatiia morfologicheskoi omonimii na osnove normalizuiushchikh podstanovok i pozitsii sosednikh slov [Probability model of morphological disambiguation based on the normalized substitutes and position of neighbouring words]. *Komp'iuternaia lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnogo seminara «Dialog-2005»*. Moscow, Nauka Publ., 2005.

Статья поступила в редакцию 1.06.2013.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Порохнин Алексей Александрович – Московский государственный технический университет им. Н. Э. Баумана; магистрант кафедры «Программное обеспечение ЭВМ и информационные технологии»; alexey.porokhnin@gmail.com.

Porokhnin Alexey Aleksandrovich – Moscow State Technical University named after N. E. Bauman; Candidate for a Master degree of the Department "Computer Software and Information Technology"; e-mail: alexey.porokhnin@gmail.com.