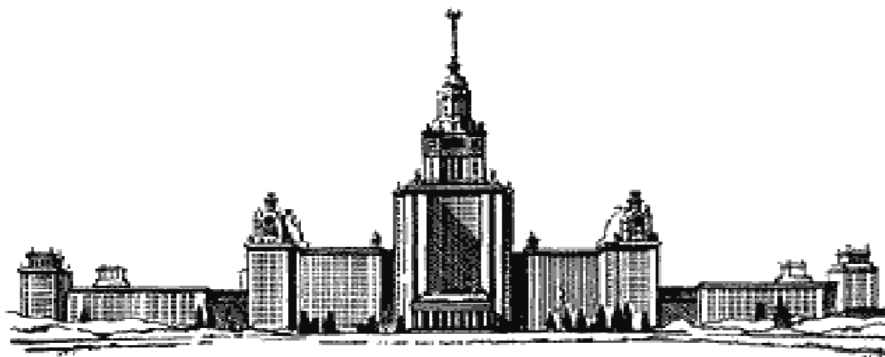


**Московский Государственный Университет
им. М.В. Ломоносова**



**Снятие морфологической омонимии на основе сокращённой
модели Маркова**

Студент: Колокольников Александр

Группа: 422

Факультет: ВМК

Преподаватель: Dr. Мстислав Масленников

Москва, ноябрь 2013г.

Аннотация

В работе описывается эксперимент по снятию морфологической неоднозначности с использованием анализатора построенного на базе сокращённой модели Маркова (CoMM).

Мы проверили гипотезу, что при реализации марковской модели достаточно использовать локальную оценку биграмм вместо полной оценки пути. В результате получился результат Accuracy=92.7%, сравнимый с оценкой лучшей из известных нам моделей.

1 Введение

В настоящее время нам неизвестны программы с открытым исходным кодом, снимающие морфологическую омонимию для русского языка.

Одна из предыдущих работ на эту тему от Сокирко и Толдовой (2004) использует скрытую модель Маркова. Однако, при её использовании необходимо оценивать пройденный путь. При этом вероятностная оценка текущего слова умножается на оценку пути, ведущего к данному слову. На наш взгляд, такой метод приводит к оперированию маленькими числами, порядок которых зависит от длины предложения. Если предложение оказывается слишком длинным, то число может оказаться нулевым.

Чтобы избежать малых чисел при оценке пути, мы применяли модификацию скрытой модели Маркова, названную сокращённой моделью Маркова (СоММ), для снятия морфологической омонимии. При этом мы исследовали, нужно ли оценивать полную оценку пройденной цепочки, или же достаточно использовать локальную оценку биграмм.

2 Методика решения

Наша система состоит из генерации словоформ и использования сокращённой модели Маркова (СоММ). Архитектура системы приведена на рис. 1 снизу.

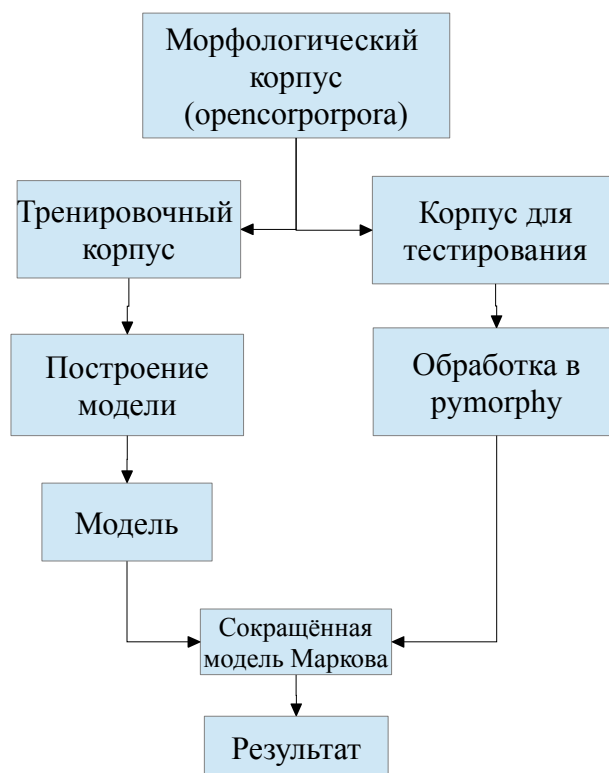


Рисунок 1. Архитектура системы СоММ для снятия омонимии

2.1 Генерация словоформ

Для генерации словоформ мы использовали морфологический анализатор Rymorphy, реализованный на языке Python. Для каждого слова текста Rymorphy строит несколько словоформ, содержащих различные характеристики слова. Например, для слова «стали» строятся словоформы, соответствующие производной от глагола «стать», и существительному от слова «сталь». Другой пример для предложения «Эти типы стали есть на складе» приведён в Таблице 1.

word	base	tag
<i>Эти</i>	этот	ADJF,Apro,plur,nomn ADJF,Apro inan,plur,accs
<i>типы</i>	тип	NOUN,inan,masc plur,accs NOUN,inan,masc plur,nomn NOUN,anim,masc plur,nomn
<i>стали</i>	стать	VERB,perf,intr plur,past,indc
	сталь	NOUN,inan,femn sing,gent NOUN,inan,femn sing,datv NOUN,inan,femn sing,loct NOUN,inan,femn plur,nomn NOUN,inan,femn plur,accs
<i>есть</i>	есть	INTJ INFN,impf,tran
	быть	VERB,impf,intr sing,3per,pres,indc VERB,impf,intr plur,3per,pres,indc,Infr
<i>на</i>	на	PREP INTJ PRCL
<i>складе</i>	склад	NOUN,inan,masc sing,loct

Таблица 1. Пример разбора предложения с помощью Rymorphy

Таким образом, Rymorphy позволяет (а) приводить слово к нормальной форме (например, «люди -> человек»); (б) ставить слово в нужную форму (например, во множественное число, менять падеж и т. д.); и (в) возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т. д.).

2.2 Сокращённая модель Маркова

На следующем шаге нам нужно выбрать правильные теги из сгенерированных программой PyMorphy. Например, для слова «эти» нам нужно выбрать тег (ADJF,Apro,plur,nomn). Для этого мы применяем сокращённую модель Маркова.

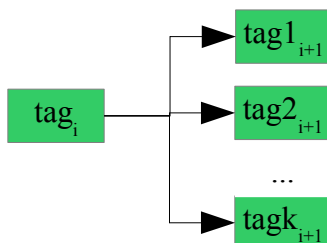


Рисунок 2. Пример последовательности тегов

Для каждого тега tag_i мы подсчитываем вероятность тега $P(tag_i)$, вероятность базы слова $P(base_i)$, и также условную вероятность $P(tag_{i+1}|tag_i)$ на корпусе для обучения (см рис. 2). Наборы этих вероятностей составляют полученную модель.

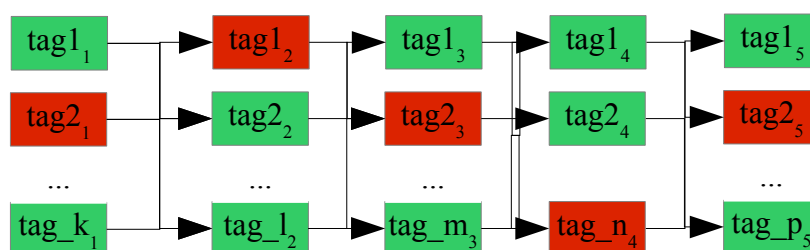


Рисунок 3. Пример последовательности тегов

Чтобы применить эту модель, мы проходим каждое предложение слева направо. На начальном шаге мы отбираем тег, соответствующий максимальной вероятности $P(base_i)$. На каждом последующем шаге мы выбираем тег с максимальной вероятностью перехода $P(tag_{i+1}|tag_i)$, как изображено на рис. 3.

3 Эксперименты

Нашей целью была проверка гипотезы, что достаточно использовать локальную оценку биграмм вместо полной оценки пути для снятия морфологической неоднозначности. Для этого мы использовали корпус оренсгорга со снятой омонимией¹. Этот подкорпус также содержит грамматическую информацию о свойствах слов. Мы разбивали подкорпус в соотношении 9:1. Знаки препинания и английские слова были проигнорированы перед обработкой.

¹ http://opencorpora.org/files/export/annot/annot.opcorpora.no_ambig.xml.zip

В результате мы получили точность Ассигасу = 92.7%. Она сравнима с точностью 94.6%, полученных в работе Сокирко и Толдовой (2004) на основе скрытой модели Маркова. Полученный результат доказывает, что достаточно использовать локальную оценку вместо полной оценки пути в скрытой модели Маркова.

Литература

Сокирко, Толдова. 2004. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп).
<http://www.aot.ru/docs/RusCorporaHMM.htm>