УДК 004.8

Математическое описание задачи построения дерева синтаксического подчинения с использованием чанкинга и словаря биграмм[1]

В. Н. Поляков
НИТУ МИСИС, Институт Языкознания РАН
pvn-65@mail.ru

Аннотация. В работе приводится описание и формальное решение математической задачи редукции матрицы потенциальных словоформ и построения дерева зависимостей. Матрица потенциальных словоформ является удобным представлением предложения на естественном языке и содержит слова и возможные морфоформы. Обычно задача парсинга ЕЯ-предложения вызывает оправданные трудности из-за высокого уровня омонимии. В статье приведены этапы редукции матрицы и множества синтаксических связей, упрощающие и ускоряющие поиск решения. Использована нотация неполного грамматического разбора (чанкинга) в грамматике зависимостей для русского и английского языков. Найденное решение применимо в приложениях разрешения омонимии, омофонии и коррекции правописания. Модель использует синтаксический контекст, включая чанкинг и словарь биграмм.

Ключевые слова: синтаксис, модель зависимостей, чанкинг, биграммы, омонимия, омофония, коррекция правописания, русский, английский.

Mathematical description of the problem of building a tree of syntactic subordination using chunking and a dictionary of bigrams [2]

V. N. Polyakov
NUST MISIS, Institute of Linguistics of RAS
pvn-65@mail.ru

Abstract. The present research draws upon the description and formal solution of the mathematical problem of reducing the matrix of potential word forms and building a dependency tree. The matrix of potential word forms is a convenient representation of a clause in a natural language, and it contains words and all possible morphological forms. As a rule, the task of parsing a clause in a natural language causes certain difficulties due to the high level of homonymy. The paper describes stages aimed at reducing the matrix and the set of syntactic connections that facilitate and hasten the search for the solution. The work uses notation of incomplete grammatical analysis (chunking) in dependency grammar for English and Russian. The suggested solution can be used for homonymy and homophony disambiguation and in spelling correction. The model is also based on syntactic contect, including chunking and a dictionary of bigrams.

Keywords: syntax, dependency model, chunking, bigrams, homonymy, homophony, spelling correction, Russian, English.

[1] Работа выполнена при поддержке гранта РНФ № 15-11-10019.
[2] The research was supported by RSF grant № 15-11-10019.

Введение

В настоящее время существуют две наиболее популярные синтаксические модели: формальная грамматика Хомского (Chomsky 1956, 1957) и грамматика зависимостей Теньера (Tesnière 1959, 1988). За 60 лет с момента начала работ этой области произошел существенный прогресс в области построения синтаксических моделей для естественного языка.

Были созданы парсеры (программы грамматического разбора) для целого ряда языков: английский, русский, французский, немецкий, турецкий, арабский, чешский, болгарский и многие другие (Gerdes et al. 2011; Hajičová et al. 2013; Hajičová and Nivre, 2015). Были созданы банки деревьев (treebanks), которые содержат синтаксические описания по типу корпусов текстов для обеих грамматик. Однако и в области генеративных грамматик, и в области грамматик зависимостей наблюдается дефицит практических приложений. И связано это, в первую очередь, со сложностью применения этих моделей на практике.

Считается, что грамматика зависимостей больше подходит для языков со свободным порядком слов, к которым относится русский, так же, как и большинство славянских языков. Также надо отметить, что грамматика зависимостей ближе по своей структуре к логическому представлению, чем генеративная грамматика.

В отличие от синтаксиса Мельчука (Melchuk 1987, 2003, Апресян и др. 1981 , Apresjan et al. 1992), где описание каждой синтаксической ситуации представлено в виде синтагм наиболее полно и детально, синтаксис чанкинга описан очень кратко. Это позволяет приступить к тестированию модели на практике, не дожидаясь полного синтаксического описания языка.

Модель неполного синтаксического разбора – чанкинга – была протестирована в рамках задачи автоматической коррекции правописания. Первая версия программы коррекции правописания для английского и русского языков была написана в 2016 (Anisimov et al. 2016-1, Anisimov et al. 2016-2) с использованием фреймфорка UIMA[3] (Unstructured Information Management Architecture - Архитектура управления неструктурированной информацией), языка программирования Java и программной библиотеки NLP@Cloud (Moscow). Именно этот фреймворк использовался в проекте IBM Watson (Ferrucci et al. 2013). Изначально фреймворк UIMA разрабатывался фирмой IBM, и затем был передан в фонд Apache Software Foundation. В дальнейшем работа над двумя вариантами программы парсинга (для русского и для английского языков) велась параллельно.

Работы над библиотекой обработки текстов на естественном языке (ОЕЯТ) NLP@Cloud были начаты в 2012 году. В настоящее время существует две версии библиотеки: NLP@Cloud(Moscow) в НИТУ «МИСИС» и NLP@Cloud(Kazan) в КФУ (Поляков и др. 2013). Версия NLP@Cloud(Kazan) ориентирована на обработку больших коллекций текстов для выявления именованных сущностей (имен собственных), именных групп и т.д., в том числе с использованием облачных вычислений. Версия NLP@Cloud(Moscow) ориентирована на десктоповские приложения для углубленной ОЕЯТ.

В круг перспективных прикладных задач библиотеки NLP@Cloud(Moscow) на основе чанкинга входят задачи разрешения омонимии, коррекции правописания и разрешения омофонии.

В данной работе выполнено математическое описание задачи коррекции правописания с использованием матрицы потенциальных словоформ. Это представление оказалось перспективным, так как задачи снятия омонимии и коррекции правописания сводятся к проблеме редукции размерности матрицы потенциальных словоформ и множества

[3] UIMA Homepage at the Apache Software Foundation https://uima.apache.org/ (Дата обращения 28.05.2017).

синтаксических связей. Формальная постановка выполнена в теоретическо-множественной модели. Результаты работы программы приведены в ссылках и в литературе.

1. Описание задачи редукции матрицы потенциальных словоформ

В данном разделе приводится описание математической задачи сокращения (редукции) матрицы потенциальных словоформ. Изложение базируется на примерах прикладных задач разрешения омонимии (пример 1) и коррекции правописания (пример 3). Ракрывается методика сокращения числа переборов, содержащая 6 этапов.

1.1. Матрица потенциальных словоформ

Будем описывать исходное предложение в виде матрицы омонимичных форм (в дальнейшем, матрица потенциальных словоформ − МаПС). МаПС подойдет нам и для описания вариантов коррекции правописания и распознанных в речи слов (в случае речевого интерфейса). Общий вид МаПС приведен на рис.1.

Элементами матрицы МаПС

$$\mathbf{MaPF(i,j)} \quad (1)$$

являются кортежи (лемма + грамматический вектор),
где:

$\mathbf{m_{ij}} = (\mathbf{lemma_{ij}}, \mathbf{gr\_vector_{ij}})$ – кортеж «лемма + грамматический вектор» для словоформы **i** слова **j**; (2)

$\mathbf{lemma_{ij}}$ – лемма **i** для слова $\mathbf{w_j}$; (3)

$\mathbf{gr\_vector_{ij}}$ – элемент множества грамматических векторов **GR_VECTOR** для леммы **ij**; (4)

Некоторые $\mathbf{m_{ij}} = \mathbf{Null}$ (нет значения)

i – индекс словоформы;

j – индекс слова.

Номер слова j / Номер морфоформы i	w_1	w_2	…	w_j	…	w_n
1	m_{11}	m_{21}	…	m_{1j}	…	m_{1n}
2	m_{21}	m_{22}	…	m_{2j}	…	m_{2n}
...	…	…	…	…	…	…
i	m_{i1}	m_{i2}	…	m_{ij}	…	m_{in}
…	…	…	…	…	…	…
r	m_{r1}	m_{ir}	…	m_{rj}	…	m_{rn}

Рис.1. Матрица потенциальных словоформ.

1.1.1. Пример омонимии

Приведем пример МаПС для простого русского предложения.

Пример 1.

В настоящее время ведущие производители в состоянии изготавливать в промышленных масштабах этот нанопорошок (2). [4]

[4] Пример взят из: http://www.ruscorpora.ru (Дата обращения: 28.05.2017).

МаПС для примера 1 приведена в Приложении 1 (тело матрицы снабжено заголовками строк и столбцов). Структура грамматического вектора не приводится, чтобы не загромождать описание. Она соответствует описанию, принятому в OpenCorpora. Для более детального ознакомления с этим вопросом рекомендуется литература (Зализняк 1980, Bocharov et al. 2011: 10-17).

1.2. Число переборов при решении задачи синтаксического разбора

В исходной МаПС количество синтаксических комбинаций очень велико, так как для построения синтаксической структуры предложения прямым перебором потребуется попытаться скомбинировать каждое слова с каждым. Сюда надо добавить связи и с ошибочными морфоформами. Это потребует организации связей всех элементов со всеми в матрице МаПС, кроме вертикальных элементов, т.к. из всех вариантов слова в предложение входит только один.

Это количество задается через число парных сочетаний $\mathbf{N^k}$ (Абрамовиц и Стиган 1979) и выражается формулой:

$$\mathbf{N^k = 2 \times N_1 - 2 \times n \times N_2}, \quad (5)$$

где:

$\mathbf{N^k}$ –общее число потенциальных синтаксических связей в матрице МаПС размером $\mathbf{n \times r}$ элементов;

$\mathbf{N_1}$ - число парных комбинаций в множестве из $\mathbf{n \times r}$ элементов;

$\mathbf{N_2}$ - число парных комбинаций в множестве из $\mathbf{r}$ элементов;

$$\mathbf{N_1 = (n \times r)!/ ((n \times r)-2)! \times 2!)} \quad (6)$$

$$\mathbf{N_2 = (r)!/ (r-2)! \times 2!)} \quad (7)$$

где:

$\mathbf{N_1}$ – число парных комбинаций в матрице размером $\mathbf{n \times r}$,

$\mathbf{N_2}$ – число парных сочетаний в вертикальных столбцах,

$\mathbf{n}$ – число слов в предложении;

$\mathbf{r}$ – уровень омонимии.

Число синтаксических связей удваивается, так как любое слово потенциально может быть и главным и зависимым.

В таблице 1 приведены данные количества переборов для предложений, содержащих число слов в диапазоне 3 ... 15, с уровнем омонимии 2 ... 13.

Таблица 1.

Число переборов в предложении

Число слов в предложении, n	**Уровень омонимии, r**	**N_1**	**N_2**	**Число комбинаций N**	**Прим.**
5	2	45	1	80	
6	2	66	1	120	
7	2	91	1	168	
3	3	28	3	54	
4	3	66	3	108	
5	3	105	3	180	
6	3	153	3	270	

Число слов в предложении, n	Уровень омонимии, r	N_1	N_2	Число комбинаций N	Прим.
7	3	210	3	378	
...	...	…	...	…	
13	13	14196	78	26364	1
15	6	4005	15	7560	2

Примечания: 1. Соответствует примеру 1 (Раздел 1.1.1., Приложение 1, табл. 1-1).
2. Соответствует примеру 3 (Раздел 2; Приложение 2, табл. 2-1).

Вычислительная сложность на этом этапе составляет O((n×r)(n×r -1)).

Кроме того, в общем случае при построении дерева синтаксического подчинения необходимо решить задачу, подобную задаче построения *минимального остовного дерева.* Для этого существует известный алгоритм Борувки/Соллена (Кормен и др. 2005). Однако он нам не подходит, так как как дерево зависимостей строится из списка чанков (готовых ребер графа), при этом на задачу накладывается ряд ограничений, а именно:

- наличие списка грамматических основ;

- чанк должен входить в БД «Чанкинг»;

- чанки соединяются между собой следующим образом: зависимая часть вышестоящего чанка становится главной частью нижестоящего;

- используется не более одного омонима для каждого слова в дереве;

- каждое слово может быть использовано в качестве зависимого в чанке только один раз.

Кроме того, грамматический словарь, представленный на сайте http://opencorpora.org/dict.php (Bocharov et al. 2011), обладает следующей особенностью: все буквы русского алфавита имеют омонимичные формы маркированные как *имя существительное*. Очевидно, что это предусмотрено для автонимного употребления знака (см. пример 2).

Пример 2:
«Буква «я» – последняя буква в алфавите.

В примере 2 буква «я» используется автонимно, но при этом ее написание совпадает с местоимением *я*. Подобная практика приводит в резкому возрастанию числа потенциальных чанков и зашумляет распознавание в предложении однобуквенных предлогов («в», «к», «о», «с», «у»), союзов («а», «и»), и местоимений («я»). Рассмотрим количество ложных чанков, возникших в результате «наведенной» омонимии, связанной с описанным явлением для предложения «*Я *попогла какому-то *мужщине тележку в *магазни *заветри*» в задаче коррекции правописания.

Всего программой был сгенерирован 4991 чанк. После очистки списка омонимов на этапе «Чистка списка омонимов и работа со словарем» общее число чанков в указанном предложении сократилось до 513 (на 81%).

1.3. Методики сокращения числа переборов, предложенные в работе

1.3.1. Чистка списка омонимов и работа с морфологическим словарем

В пайплайне был предусмотрен этап «Чистка списка омонимов и работа с морфологическим словарем». В рамках этого этапа выполняется следующее: программа

исключает «ложные» омонимы (речь идет о случаях совпадения однобуквенных предлогов, союзов и местоимений с буквами алфавита). Кроме того, ввиду чувствительности словаря к букве «ё», на этом этапе исправляются слова, содержащие ее.

1.3.2. Применение словаря биграмм

Данный этап состоит в исключении наименее вероятных вариантов коррекции с помощью словаря биграмм (Davies 2009, 2013, Bocharov et al. 2001). Для каждого варианта коррекции, сгенерированного на этапе «Синтез потенциальных коррекций», строится две биграммы – со словом, стоящим слева, и со словом, стоящим справа. В случае, если ошибочное слово является первым или последним словом в предложении, программа строит только одну биграмму.

Если обе биграммы присутствуют в словаре, вариант коррекции помечается как «очень ожидаемый». Если в словаре находится только одна биграмма из двух, вариант коррекции помечается как «малоожидаемый». Если же ни одна биграмма не была найдена в словаре, вариант коррекции помечается как «неожидаемый». Для дальнейшей работы оставляются только варианты коррекции с наиболее высокой степенью уверенности (ожидания) (т.е., при наличии «очень ожидаемых» вариантов коррекции «малоожидаемые» и «неожидаемые» исключается из списка и в дальнейшем не рассматриваются; при наличии только «малоожидаемых» и «неожидаемых», оставляются только «малоожидаемые»).

Описанный алгоритм оказался эффективным способом исключения части неверных коррекций. В некоторых случая после выполнения этапа «Фильтрация коррекций с помощью словаря биграмм» для слова остается только один вариант замены. В других случаях исключается часть сгенерированных коррекций, что упрощает и ускоряет дальнейшую работу и повышает вероятность выбора нужной замены с помощью чанкинга.

Интересно отметить, что в русском и английском языках словари биграмм работают по-разному. Для русского языка словарь биграмм не имеет разметки ни по грамматическому вектору, ни по частям речи. Из-за этого применение словаря биграмм для редукции матрицы МаПС ограничивается задачами коррекции правописания и разрешения омофонии (выбор из списка одинаково звучащих слов). Это связано с тем, что словарь русских биграмм является «слепым» относительно омонимии. Словарь английских биграмм, напротив, имеет разметку по частям речи (PoS-tagging), благодаря чему он работает и в задаче разрешения омонимии корректных предложений, и в задаче коррекции правописания, и в задаче разрешения омофонии.

1.3.3. Эвристика чанкинга, сокращающие число элементов матрицы МаПС

Внутри каждого предложения ищутся: а) предлоги; б) частицы «не», «ни»; в) частицы «бы»; г) наречия – интенсификаторы действия и качества. Для каждого предлога, частицы и наречия ищется связанное слово, например: «в ресторане», «после бала», «не он», «хотелось бы». Наличие у слова предлога, частицы, наречия отмечается в расширенном грамматическом векторе слова, при этом сами слова исключаются из матрицы МаПС.

Таким образом, чанк становится двухчастным, т.е. состоящим из двух частей (главная часть – зависимая часть). Части чанка, которые носят синтаксический характер, но выражены лексически (предлог как маркер падежа, союз, частицы «не», «бы», наречия-интенсификаторы качества (т.е. прилагательного) и действия (т.е глагола)) выносятся из лексического состава предложения (матрицы МаПС и в последующем – дерева подчинения) в грамматическую структуру. Этим достигается однородность синтаксической структуры дерева чанков и алгоритмичность его обработки.

1.3.4. Использование БД «Чанкинг»

Строится список потенциальных чанков путем комбинации всех словоформ со всеми, за исключением вертикальных столбцов, элементы которых не комбинируются между собой. Потенциальные чанки, которые не обнаружены в базе данных (БД) «Чанкинг», исключаются из рассмотрения. БД «Чанкинг» была разработана специально для данного проекта (русский и английский вариант) (Anisimov et al. 2016-1, Anisimov et al. 2016-2) и является уникальной.

1.3.5. Выбор грамматической основы

На основании схемы грамматических основ, разработанной в рамках данного проекта (Anisimov et al. 2016-1, Anisimov et al. 2016-2) находятся все потенциальные грамматические основы предложения, которые в дальнейшем включают вершины синтаксических деревьев. Таким образом, исключается необходимость строить **n** × **r** деревьев. Были разработаны две схемы для выбора грамматических основ – для русского и для английского языков.

1.3.6. Построение дерева зависимостей

Каждый чанк, обнаруженный в схеме грамматических основ, включает вершину дерева. Далее требуется построить дерево чанков на простом предложении (клаузе). Один чанк считается соединенным с другим, если главное слово первого является зависимым словом второго. Граф чанков обходится в ширину, при этом отсекаются все «короткие» ветви (такие, которые не имеют продолжения, при условии, что хоть одна из текущих ветвей-потомков продолжение имеет) по пути. Строим такое дерево для всех возможных пар подлежащее-сказуемое, выделенных в клаузе (Зыков 2004).

По мере построения деревьев каждое новое дерево сравнивается с текущим лучшим деревом по количеству слов предложения, включенных в него. Если новое дерево включает в себя больше слов исходного предложения, оно становится лучшим, и следующее дерево сравнивается уже с ним.

Варианты коррекции слов с орфографическими ошибками, включенные в лучшее дерево, считаются верными.

2. Иллюстрация работы методик сокращения числа переборов

Рассмотрим работу методик сокращения числа переборов на примере 3.

Пример 3.
*Что вам *нравиься в *ваших телефонах.*

Пример построен на итогах работы пайплайна исправления ошибок правописания. В приложении 2 (Табл. 2-1) приведена матрица МаПС, где собраны все омонимы правильно введенных слов и омонимы для всех сгенерированных вариантов замены для ошибочных слов. В приложении 2 (Табл. 2-2) отображена легенда, показывающая на каком этапе была проведена редукция.

В результате чанкинга было отобрано 18 чанков (потенциальных ребер), представленных в Приложении 3 (словоформы, т.е. части чанков, выделены жирным).

В результате выбора грамматических основ было отобрано три чанка, включающие вершины потенциальных деревьев зависимостей. Они представлены ниже (словоформы выделены жирным):

```
Head chunks: 3
```

1. {chunkId=1, template={id=231},
   mainWord={text='Что',correction='**что**',wordform=[pos: "NPRO" lemma: "что"
   grammems: ["neut", "sing", "nomn"]],wordInd=0},
   dependentWord={text='нравиься',correction='**нравится**',wordform=[pos:
   "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per",
   "pres", "indc"]],wordInd=2}, isHeadChunk=true}
2. {chunkId=2, template={id=233},
   mainWord={text='Что',correction='**что**',wordform=[pos: "NPRO" lemma: "что"
   grammems: ["neut", "sing", "nomn"]],wordInd=0},
   dependentWord={text='вашх',correction='**ваша**',wordform=[pos: "ADJF" lemma:
   "ваш" grammems: ["femn", "sing", "nomn", "Apro"]],wordInd=4},
   isHeadChunk=true}
3. {chunkId=3, template={id=37},
   mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB"
   lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres",
   "indc"]],wordInd=2},
   dependentWord={text='**телефонах**',correction='телефонах',wordform=[pos:
   "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]
   ],wordInd=5}, isHeadChunk=true}

В результате редукции МаПС и работы программы было построено дерево синтаксического подчинения (рис. 2).

```
Best tree:
[что] [нравится] (templateId: 231, chunkId: 1)
  [нравится] [телефонах] (templateId: 37, chunkId: 3)
    [телефонах] [ваших] (templateId: 1, chunkId: 15)
```

Рис. 2. Дерево синтаксического подчинения, построенное в результате редукции МаПС

Дерево представлено на рис. 3 в привычном для грамматик зависимостей формате.

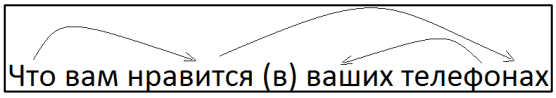

Рис. 3. Дерево для примера 3

Хочется обратить внимание на то, что в дереве (рис. 2) не нашла отражение одна синтаксическая связь («нравится вам»). Это объясняется неполнотой описания всех чанков в БД «Чанкинг». Тем не менее, несмотря на этот недостаток, программа редукции работает и корректирует правописание. В программе была использована БД «Чанкинг» (русский) (версия: «весна, 2016»), которая содержит описание 242 чанков русского языка[5].

3. Формальная постановка задачи

В разделе 3 приведено формальное описание задачи редукции матрицы потенциальных словоформ и построения лучшего дерева синтаксического подчинения. Постановка задачи выполнена в логико-математической нотации с использованием теории множеств, теории отношений, теории матриц и теории графов (Зыков 2004). Для решения задачи была введена новая операция грамматической дизъюнкции множеств. Результаты работы программы по проверке модели даны в ссылках и в литературе.

[5] БД Чанкинг (английский) (версия: весна, 2016), содержит описание 228 чанков английского языка.

Дано.

Предложение, содержащее омонимичные формы, а также варианты коррекции правописания или распознанные в речи слова.

Задача.

Найти наилучшее представление данного предложения в виде однозначных кортежей морфоформ и построить дерево зависимостей.

Решение.

Опишем исходное предложение в виде матрицы потенциальных словоформ МаПС:

$$\mathbf{MaPF(i,j),\ i = 1 \ldots r,\ j = 1 \ldots n.} \quad (8)$$

Последовательно применяем этапы сокращения (редукции) размерности матрицы **MaPF(i,j).**[6]

Этап 1. Снятие ложной омонимии.

Сокращаем количество омонимов за счет снятия ложной омонимии.

$$\mathbf{M\_REDUC = M\_ORIG \setminus M\_FALSE\_HOMON,} \quad (9)$$

где:
M_ORIG – исходное множество омонимичных форм до снятия ложной омонимии;
M_FALSE_HOMON – множество «ложных» омонимов;
M_REDUC – редуцированное множество омонимичных форм после снятия «ложной» омонимии;
A \ B – означает разность множеств **A** и **B**.

Этап 2. Применение словаря биграмм.

Сокращение число омонимов за счет применение словаря биграмм.

$$\mathbf{BIGRAM\_REDUC = BIGRAM\_ORIG \cap BIGRAM,} \quad (10)$$

где:
BIGRAM – словарь биграмм;
BIGRAM_ORIG – исходное множество биграмм в предложении, полученное соединением пар соседних слов;
BIGRAM_REDUC – отфильтрованное (редуцированное) множество биграмм в предложении, полученное пересечением двух множеств.

Этап 3. Применение эвристик чанкинга.

Сокращаем число слов, а следовательно, размерность МаПС, за счет реализации эвристик чанкинга, переводящих слова, которые носят синтаксический характер, но выражены лексически, в кортеж связанных с ними слов.

[6] Множества будем обозначать с прописной (большой) буквы (пример: **MaPF, M_REDUC**). Элементы множеств будем обозначать строчными (малыми) буквами (пример: $\mathbf{lemma_{ij}}$, $\mathbf{gr\_vector_{ij}}$, $\mathbf{extension_{ij}}$). Переменные в формулах будем выделять жирным шрифтом. Будем использовать подстрочные знаки для обозначения индексов (пример: $\mathbf{h_{ji\ op}}$), надстрочные знаки (пример: **R*** – для обозначения преобразованных множества.

$$\mathbf{W\_REDUC = W\_ ORIG \setminus W\_SHIFTED,} \quad (11)$$

где:

W_ ORIG – начальное множество слов в предложении;

W_SHIFTED – множество слов (предлоги, союзы, частицы «не», «бы», наречия-интенсификаторы качества (т.е. прилагательного) и действия), перемещенные в кортеж грамматического вектора связанного слова;

W_REDUC – множество слов в предложении после реализации эвристик чанкинга.

Таким образом, количество омонимов еще сокращается. Получаем матрицу МаПС_РЕДУК сокращенной размерности

$$\mathbf{MaPF\_REDUC(i,j),\ i = 1 \ldots r\_reduc,\ j = 1 \ldots n\_reduc}. \quad (12)$$

Элементами матрицы МаПС_РЕДУК являются кортежи расширенного грамматического вектора (лемма + грамматический вектор + расширение),

где:

$\mathbf{m\_ext_{ij} = (lemma_{ij}, gr\_vector_{ij}, extension_{ij})};$ (13)

$\mathbf{lemma_{ij}}$ – лемма **i** для слова $\mathbf{w_j}$; (14)

$\mathbf{gr\_vector_{ij}}$– грамматический вектор для леммы **ij**; (15)

$\mathbf{extension_{ij}}$ – расширение грамматического вектора. (16)

Некоторые $\mathbf{m\_ext_{ij}}$ = **Null** (нет значения);

$\mathbf{m\_ext_{ij} \in M\_EXT;}$

i , j – индексы словоформ и слов соответственно;

r_reduc, n_reduc – число словоформ и слов в матрице **MaPF_REDUC** соответственно;

Отметим, что термин *расширение грамматического вектора* означает обогащение описания словоформы за счет включения в ее кортеж информации о служебных частях речи.

В тоже время термин *редукция* применятся в данной статье для описания процесса сокращения размерности матрицы потенциальных словоформ. В некоторой степени эти процессы связаны между собой, так как редукция размерности матрицы производится, в том числе, и за счет расширения грамматического вектора.

Этап 4. Использование базы данных «Чанкинг».

Перейдем к графовому представлению. Построим граф $\mathbf{G^0}$, включающий все возможные деревья зависимостей:

$$G^0 (M\_EXT^{\Sigma}, H^{\Sigma}), \quad (17)$$

где:

M_EXT$^{\Sigma}$– множество узлов графа, т.е. словоформ предложения, обработанное на этапах 1, 2, 3 и представленное в МаПС_РЕДУК (формула (12)). Элементами множества **M_EXT**$^{\Sigma}$ являются значения расширенного грамматического кортежа $\mathbf{m\_ext_{ji}}$ (формула (13)).

H$_{\Sigma}$ – множество ребер начального графа, составленное из максимального количества потенциальных синтаксических связей, составленных на множестве **M_EXT**$^{\Sigma}$**:**

$$H^{\Sigma} = \{(m\_ext_{ji}, m\_ext_{op})\}, \text{ для } \forall: , i,o \in R^*, j \neq o, j,p \in N^*, \quad (18)$$

где:

$\mathbf{m\_ext_{ji}}$, $\mathbf{m\_ext_{po}}$ – кортежи расширенного грамматического вектора;

$\mathbf{m\_ext_{ji}}$, $\mathbf{m\_ext_{po}} \in \mathbf{M\_EXT}^{\Sigma}$ – они принадлежат множеству кортежей расширенного грамматического вектора; (19)

R*, N* – редуцированные множества индексов для словосочетаний и слов соответственно;

i,o – индексы словоформ;

j,p – индексы слов.

Введем для **H**$_{\Sigma}$ сокращенное обозначение:

$$H^{\Sigma} = \{(m\_ext_{ji}, m\_ext_{op})\} = \{H_{ji\ op}\}, \quad (20)$$

где:

$\mathbf{h_{ji\ op}} \in \mathbf{H}^{\Sigma}$.

Число потенциальных синтаксических связей равно числу парных комбинаций всех элементов (кортежей расширенного грамматического вектора) матрицы МаПС_РЕДУК, умноженному на два. кроме вертикальных. Удвоение означает, что каждый элемент матрицы потенциально может служить и опорной, и зависимой частью чанка. Кроме того, запрет на синтаксические связи в столбцах означает, что разные словоформы одного слова не могут образовывать синтаксические связи между собой.

База данных «Чанкинг» содержит все разрешенные в языке комбинации типов словосочетаний. БД «Чанкинг» составлена экспертами-лингвистами.

Использование БД «Чанкинг» фактически подразумевает редукцию множества синтаксических связей за счет исключения из множества **H**$^{\Sigma}$ неописанных в БД «Чанкинг» связей.

Множество «Чанкинг» **DB_CHUN** содержит обобщенные грамматическое описания синтаксических связей (=чанков), т.е. таких парных комбинаций грамматических векторов, которые встречаются в текстах.

$$DB\_CHUN = \{(gr\_vector^{main}, gr\_vector^{depen})\}, \quad (21)$$

где:

GR_VECTOR – множество грамматических векторов (формула 4);

$\mathbf{gr\_vector^{main}}$. – грамматический вектор для опорной части чанка;

$\mathbf{gr\_vector^{depen}}$ – грамматический вектор для зависимой части чанка.

Редукция множество ребер начального графа **H**$^{\Sigma}$ происходит следующим образом.

Элементы множества отношений **DB_CHUN** (ф. 20) достраивается «пустыми» леммами, образуя таким образом $\mathbf{H^{db\_chun}}$**–** множество разрешенных ребер для графа $\mathbf{G^0}$ (см. ф. 17).

$$\mathbf{H^{db\_chun} = \{((null, gr\_vector^{main}), (null, gr\_vector^{depen}))\}.} \quad (22)$$

Назовем кортеж **(null, gr_vector)** *грамматической леммой* **gm_ext**.

Тогда:

$$\mathbf{H^{db\_chun} = \{gm\_ext^{main}, gm\_ext^{depen}\},} \quad (23)$$

где:

$\mathbf{h_f \in H^{db\_chun}}$, **для** $\mathbf{\forall: f \in F;}$

f – индекс для записей БД «Чанкинг»;

F – множество индексов для записей БД «Чанкинг».

Важно, что $\mathbf{H^{db\_chun}}$ и $\mathbf{H^{\Sigma}}$ эквивалентны по своей структуре, и, следовательно, их можно сравнивать между собой и выполнять на ними некоторые математические операции.

Введем операцию *грамматической дизъюнкции* множеств $\mathbf{H^{\Sigma}}$ и $\mathbf{H_{db\_chun}}$, основываясь на дизъюнкции значений их грамматических векторов по основанию $\mathbf{H^{\Sigma}}$:

$$\mathbf{H\_REDUC = H^{\Sigma} \bigcap_{GR\_VECTOR}^{H^{\Sigma}} H^{db\_chun}} \quad (24)$$

Это означает, что сравниваться будут грамматические векторы, входящие в кортежи элементов этих двух множеств, и, в случае их совпадения, будет оставаться элемент множества $\mathbf{H^{\Sigma}}$.

Сначала дадим неформальное описание операции грамматической дизъюнкции (Рис.5).

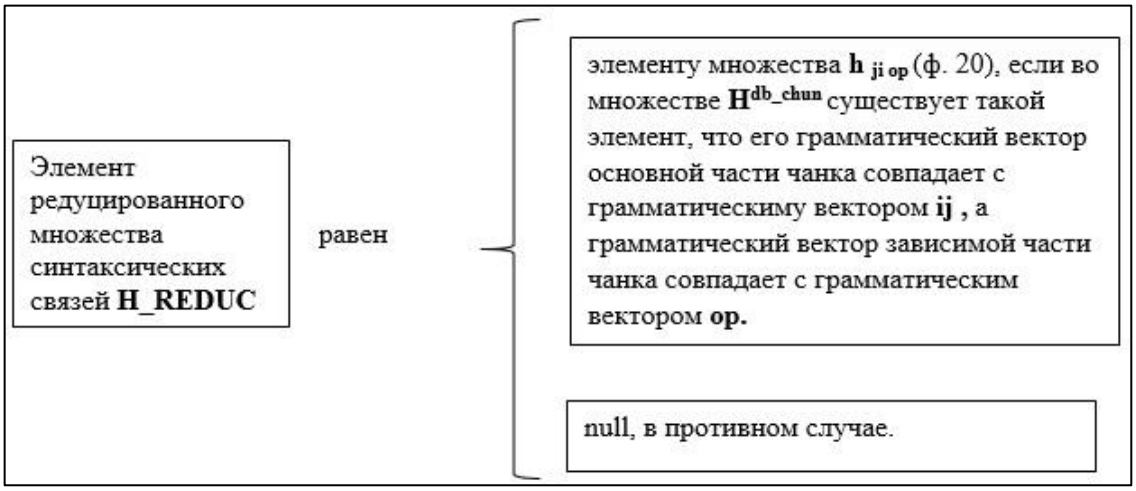

Рис. 5. Неформальное описание операции грамматической дизъюнкции

Теперь покажем, как будет выглядеть формальное описание операции грамматической дизъюнкции.

$$\mathbf{h\_reduc}_{\mathbf{ji\ op}} = \begin{cases} \mathbf{h}_{\mathbf{ji\ op}}, \text{если } \exists\,(\mathbf{gr\_vector}^{\mathbf{main}} = \mathbf{gr\_vector}_{\mathbf{ij}}) \wedge (\mathbf{gr\_vector}^{\mathbf{depen}} = \mathbf{gr\_vector}_{\mathbf{op}}) \\ \\ \text{null, если } \neg\exists\,(\mathbf{gr\_vector}^{\mathbf{main}} = \mathbf{gr\_vector}_{\mathbf{ij}}) \wedge (\mathbf{gr\_vector}^{\mathbf{depen}} = \mathbf{gr\_vector}_{\mathbf{op}}), \end{cases} \quad (25)$$

для ∀: , i,o ϵ R*, j ≠ o, j,p ϵ N*.

В результате использования БД «Чанкинг» образуется редуцированное множество ребер:

$$\mathbf{H\_REDUC} = \{\mathbf{h\_reduc}_{\mathbf{ji\ op}}\} \quad (26)$$

Это редуцированное множество ребер образует новый (редуцированный) граф **G_REDUC**:

$$\mathbf{G\_REDUC}(\mathbf{M\_EXT}^{\Sigma}, \mathbf{H\_REDUC}), \quad (27)$$

Этап 5. Выбор грамматических основ.

Изначально нам дано множество грамматических основ $\mathbf{H}^{\mathbf{s\_p}}$, которое является подмножеством множества разрешенных ребер для графа (ф. 27). Каждая грамматическая основа потенциально может содержать вершину дерева синтаксического подчинения. Множество грамматических основ сформировано экспертами-лингвистами.

$$\mathbf{H}^{\mathbf{s\_p}} \epsilon\ \mathbf{H}^{\mathbf{db\_chun}}, \quad (28)$$

где:

$\mathbf{H}^{\mathbf{s\_p}} \epsilon\ \mathbf{H}^{\mathbf{db\_chun}}$ – множество грамматических основ является подмножеством множества разрешенных ребер для графа (чанков).

Результаты программных экспериментов показали, что выбор нескольких грамматических основ является более выгодной стратегией, чем выбор одной, даже обладающей максимальным рейтингом. Это можно объяснить сложностью задачи парсинга вообще и неполнотой БД «Чанкинг» в частности.

Для решения подзадачи выбора грамматических основ воспользуемся операцией грамматической дизъюнкции (ф. 24-25).

$$\mathbf{H\_REDUC}^{\mathbf{s\_p}} = \mathbf{H\_REDUC} \bigcap\nolimits_{\mathbf{GR\_VECTOR}}^{\mathbf{H\_REDUC}} \mathbf{H}^{\mathbf{s\_p}} \quad (29)$$

Этап 6. Построение дерева зависимостей.

Требуется построить дерево чанков на простом предложении (клаузе). Один чанк считается соединенным с другим, если главное слово первого является зависимым словом второго. Граф чанков обходится в ширину, при этом отсекаются все «короткие» ветви (такие, которые не имеют продолжения, при условии, что хоть одна из текущих ветвей-потомков продолжение имеет) по пути. Строим такое дерево для всех возможных пар подлежащее-сказуемое (грамматических основ), выделенных в клаузе.

По мере построения деревьев, каждое новое дерево сравнивается с текущим лучшим деревом по количеству слов предложения, включенных в него. Если новое дерево включает в себя больше слов исходного предложения, оно становится лучшим, и следующее дерево сравнивается уже с ним.

Верными считаются те варианты коррекции (из матрицы МаПС (ф.1)), которые были включены в лучшее дерево в результате выполнения этапов 1-6. [7]

Из полученного леса деревьев выбираем финальный вариант графа дерева подчинения

$$G^{FINAL}(M^{FINAL}, H^{FINAL}), \quad (30)$$

где:

G^{FINAL} . – финальный вариант графа для дерева грамматических зависимостей;

M^{FINAL} . – финальный вариант множества словоформ, отобранных на этапе 6;

H^{FINAL} . – финальный вариант множества чанков (ребер графа), отобранных на этапе 6.

4. Обсуждение

Математическое описание задачи коррекции правописания на основе чанкинга, выполненное в данной работе, использует матрицу потенциальных словоформ. Такое представление является перспективным, так как проблема редукции размерности матрицы потенциальных словоформ и дерева синтаксического подчинения оказывается важной в прикладных задачах избытка словоформ. К ним относятся задачи разрешения омонимии, разрешение омофонии (при использовании голосовых помощников), коррекции правописания.

К сожалению, преимущества метода не могут быть полностью продемонстрированы, так как в работе использован словарь биграмм русского языка с неснятой омонимией.

В словаре биграмм английского языка, напротив, выполнена разметка по частям речи. Этим можно объяснить более высокие результаты работы программы коррекции правописания для английского языка.

Предложенная модель ориентирована пока только на простые предложения.

В перспективе планируется сделать модель и программу (пайплайн) где будет реализован этап сегментации, т.е. разделение сложных предложений на клаузы.

Также в дальнейшем планируется провести ряд исследований, посвященных формальному описанию модели применения словаря биграмм и формальному описанию алгоритма построения леса деревьев и выбора лучшего дерева зависимостей. Предполагается провести исследование свойств алгоритма.

5. Заключение

В работе было выполнено математическое описание задачи коррекции правописания с использованием матрицы потенциальных словоформ. Была продемонстрирована перспективность такого представления сразу для трех прикладных задач, так как задачи снятия омонимии, коррекции правописания и снятия омофонии сводятся к проблеме редукции размерности матрицы потенциальных словоформ и построения финального дерева синтаксического подчинения. Формальная постановка выполнена в теоретико-множественной модели.

Правомерность предлагаемого подхода подтверждена примерами и результатами работы программы чанкинга.

[7] Так как этап построение дерева зависимостей выполняется программной процедурой, его подробное описание не представляется возможным в рамках данной работы. Приведем ссылки на логи с результатами работы программы для русского и английского языков: https://cloud.mail.ru/public/LS2n/Q8uYCAy9c (Дата обращения: 28.05.2017).

Благодарность

Автор выражает благодарность: В.Д. Соловьеву за идею использования словаря биграмм, Е.А. Макаровой за помощь в разработке БД «Чанкинг», схем грамматических основ, списка чанков (для русского и английского), И.С. Анисимову за программное тестирование новой модели на прикладной задаче коррекции правописания.

Литература

Anisimov, I., Makarova, E., Polyakov, V. 2016-1. Chunking in dependency model and spelling correction in Russian. In Proceedings DTGS-2016, 23-24 June, St. Petersburg, Russia. Communications in Computer and Information Science series. Springer. V. 674. Pp. 565-575. DOI: 10.1007/978-3-319-49700-6_56

Anisimov, I., Makarova, E., Polyakov, V. 2016-2. Chunking in dependency model and spelling correction in Russian and English. In Proceedings 2016 SAI Intelligent Systems Conference (IntelliSys), 21-22 September 2016, London, United Kingdom. Pp. 143-150.ISBN (IEEEXplore): 978-1-5090-1121-6. ISBN (USB) - 978-1-5090-1665-5. DOI 978-1-5090-1121-6/16.

Apresjan Ju.D., Boguslavsky I.M., Iomdin L.L. et al. 1992. Systeme de traduction automatique ETAP // P.Bouillon et A.Clas (eds.), Etudes et recherches en traduction automatique. Sillery et Montreal, Presse de l'Universite de Quebec.

Bocharov, V., Bichineva, S., Granovsky, D., Ostapuk, N., & Stepanova, M. 2011. Quality assurance tools in the OpenCorpora project. In Computational Linguistics and Intelligent Technology: Proceeding of the International Conference “Dialog”, pp. 10-17.

Chomsky N. 1956. Three Models for Description of Language. // IRE Trans. Informat. Theory, 1956, v. IT-2, p. 113-124.

Chomsky N. 1957. Syntactic Structures. — The Hague: Mouton, 1957. (Reprint: Chomsky N. Syntactic Structures. — De Gruyter Mouton, 2002. — ISBN 3-11-017279-8).

Davies, M. 2009. The 385+ million word Corpus of Contemporary American English: design, architecture, and linguistic insights. Int. J. Corpus Linguist., 14 (2009), pp. 159–190

Davies, M. 2013. "Google Scholar vs. COCA: two very different approaches to examining academic English”. Journal of English for Academic Purposes 12: 155-165.

Ferrucci, David; Levas, Anthony; Bagchi, Sugato; Gondek, David; Mueller, Erik T. 2013-06-01. "Watson: Beyond Jeopardy!". Artificial Intelligence. 199: 93–105. doi:10.1016/j.artint.2012.06.009

Gerdes, K., Hajičová, E, Wanner, L. (eds). 2011. Proceedings of the First International Conference on Dependency Linguistics (Depling-2011). Barcelona, Spain. ISBN 978-84-615-1834-0

Hajičová, E, Gerdes, K., Wanner, L. (eds). 2013. Proceedings of the Second International Conference on Dependency Linguistics (DepLing-2013). Prague, Czech Republic. ISBN 978-80-7378-240-5

Hajičová, E, Nivre, J. (eds). 2015. Proceedings of the Third International Conference on Dependency Linguistics (DepLing-2015). Uppsala, Sweden. ISBN 978-91-637-8965-6

Melchuk I. & N. Pertsov. Surface syntax of English: A formal model within the Meaning-Text framework. Amsterdam; Philadelphia: Benjamins. ISBN 90-272-1515-4 (1987)

Melchuk, I. 2003. Levels of dependency in linguistic description: Concepts and problems. In Ágel et al., 170–187.

Tesnière, L. 1959. Elements of Structural Syntax (Éléments de syntaxe structural), Klincksieck, Paris. Préface by Jean Fourquet, professeur à la Sorbonne. Second edition, reviewed and corrected.

ISBN 2-252-02620-0. Re-edition of: Tesniere, L. (1959). Éléments de syntaxe structurale, Klincksieck, Paris. ISBN 2-252-01861-5

Tesnière, L. 1988. Dependency Syntax : Theory and Practice, Albany, N.Y.: SUNY Press, 1988. 428 pp.

Абрамовиц М., Стиган И. (ред.). 1979. Справочник по специальным функциям с формулами, графиками и математическими таблицами. М.: Наука — 832 с.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. 1981. Лингвистическое обеспечение в системе автоматического перевода ЭТАП-1. // Разработка формальной модели естественного языка. Новосибирск.

Зализняк А.А. 1980. Грамматический словарь русского языка. Словоизменение. М.: Русский язык – 880 с.

Зыков А.А. Основы теории графов. 2004. М.: Вузовская книга - 664 с. - ISBN: 5-9502-0057-8

Кормен, Т., Лейзерсон, Ч., Ривест, Р., Штайн, К. 2005. Глава 23. Минимальные остовные деревья // Алгоритмы: построение и анализ = Introduction to Algorithms / Под ред. И. В. Красикова. — 2-е изд. — М.: Вильямс. — 1296 с. — ISBN 5-8459-0857-4.

Поляков В.Н., Соловьев В.Д., Анисимов И.С., Пономарев А.Д. 2013. Создание нового поколения интеллектуальных систем семантической обработки текста. Нейрокомпьютеры: разработка, применение. № 1. стр. 31-39. ISSN 1999-8554.

Приложение 1. МаПС для примера 1.

Таблица 1-1. МаПС для примера 1.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13
1	(в (пред.))	(в (сущ., ед.ч., им.п.)	(в (сущ., ед.ч., р.п.)	(в (сущ., ед.ч., д.п.)	(в (сущ., ед.ч., в.п.)	(в (пред.))	(в (сущ., ед.ч., им.п.)	(в (сущ., ед.ч., р.п.)	(в (сущ., ед.ч., д.п.)	(в (сущ., ед.ч., в.п.)	(в (пред.))	(в (сущ., ед.ч., им.п.)	(в (сущ., ед.ч., р.п.)	(в (сущ., ед.ч., д.п.)	(в (сущ., ед.ч., в.п.)
в	(настоящее (прил.,ед.ч.,ср.р., им.п.))	(настоящее (прил.,ед.ч.,ср.р., в.п.))	(настоящее, (сущ.,ед.ч.,ср.р., им.п.))	(настоящее (сущ., ед.ч.,ср.р., в.п.))	null	null	null	null	null	null	null	null	null	null	null
настоящее	(время (сущ., ед.ч., ср.р., им.п.))	(время (сущ., ед.ч., ср.р., в.п.))	null	null	null	null	null	null	null	null	null	null	null	null	null
время	(ведущие (сущ., мн.ч., м.р., им.п.))	(ведущие (сущ., мн.ч., ж.р., им.п.))	(ведущие (прил., мн.ч., им.п.))	(ведущие (прил., мн.ч., в.п.))	(ведущие (прич., мн.ч., им.п.))	null	(в (сущ., ед.ч., им.п.)	(в (сущ., ед.ч., р.п.)	(в (сущ., ед.ч., д.п.)	(в (сущ., ед.ч., в.п.)	null	null	null	null	null
ведущие	(производители (сущ., мн.ч., м.р., им.п.))	null	null	null	null	(в (пред.))	null	null	(в (сущ., ед.ч., им.п.)	(промышленных (прил., мн.ч., в.п.))	null	(этот (мест., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	null	null
производители	(в (пред.))	(в (сущ., ед.ч., им.п.)	(в (сущ., ед.ч., р.п.)	(в (сущ., ед.ч., д.п.)	(в (сущ., ед.ч., в.п.)	(в (сущ., ед.ч., им.п.)	null	null	(в (сущ., ед.ч., им.п.)	(промышленных (прил., мн.ч., п.п.))	null	null	(нанопорошок, (сущ., ед.ч., м.р., в.п.))	null	null
в	(состоянии (сущ., е.д., ср.р., п.п.)	null	null	null	null	null	(состоянии (сущ., е.д., ср.р., п.п.)	(изготавливать (глаг., инф.))	(в (пред.))	(промышленных (прил., мн.ч., р.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	(этот (мест., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	null	null
состоянии	(изготавливать (глаг., инф.))	null	null	null	null	null	(изготавливать (глаг., инф.))	(изготавливать (глаг., инф.))	(в (пред.))	(промышленных (прил., мн.ч., р.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	(этот (мест., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	null	null
изготавливать	(в (пред.))	null	null	null	null	(в (пред.))	(в (пред.))	(в (пред.))	(в (пред.))	(промышленных (прил., мн.ч., р.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	(этот (мест., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	null	null
в	(промышленных (прил., мн.ч., р.п.))	(промышленных (прил., мн.ч., в.п.))	(промышленных (прил., мн.ч., п.п.))	null	null	(промышленных (прил., мн.ч., р.п.))	(промышленных (прил., мн.ч., р.п.))	(промышленных (прил., мн.ч., р.п.))	(промышленных (прил., мн.ч., р.п.))	(промышленных (прил., мн.ч., р.п.))	null	null	null	null	null
промышленных	(масштабах (сущ., мн.ч., м.р., п.п.))	null	null	null	null	(масштабах (сущ., мн.ч., м.р., п.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	(масштабах (сущ., мн.ч., м.р., п.п.))	null	null	null	null
масштабах	(этот (мест., ед.ч., м.р., им.п.))	(этот (мест., ед.ч., м.р., в.п.))	null	null	null	(этот (мест., ед.ч., м.р., им.п.))	(этот (мест., ед.ч., м.р., им.п.))	(этот (мест., ед.ч., м.р., им.п.))	(этот (мест., ед.ч., м.р., им.п.))	(этот (мест., ед.ч., м.р., им.п.))	null	(этот (мест., ед.ч., м.р., им.п.))	null	null	null
этот	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., в.п.))	null	null	null	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	(нанопорошок, (сущ., ед.ч., м.р., им.п.))	null	null	null	null
нанопорошок	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null

	в	настоящее	время	ведущие	производители	в	состоянии	изготавливать	в	промышленных	масштабах	этот	нанопорошок
i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13
6	(в (сущ., ед.ч., тв.п.)	null	null	(ведущие (прич., мн.ч., в.п.))	null	(в (сущ., ед.ч., тв.п.)	null	null	(в (сущ., ед.ч., тв.п.)	null	null	null	null
7	(в (сущ., ед.ч., п.п.)	null	null	null	null	(в (сущ., ед.ч., п.п.)	null	null	(в (сущ., ед.ч., п.п.)	null	null	null	null
8	(в (сущ., мн.ч., им.п.)	null	null	null	null	(в (сущ., мн.ч., им.п.)	null	null	(в (сущ., мн.ч., им.п.)	null	null	null	null
9	(в (сущ., мн.ч., р.п.)	null	null	null	null	(в (сущ., мн.ч., р.п.)	null	null	(в (сущ., мн.ч., р.п.)	null	null	null	null
10	(в (сущ., мн.ч., д.п.)	null	null	null	null	(в (сущ., мн.ч., д.п.)	null	null	(в (сущ., мн.ч., д.п.)	null	null	null	null
11	(в (сущ., мн.ч., в.п.)	null	null	null	null	(в (сущ., мн.ч., в.п.)	null	null	(в (сущ., мн.ч., в.п.)	null	null	null	null
12	(в (сущ., мн.ч., тв.п.)	null	null	null	null	(в (сущ., мн.ч., тв.п.)	null	null	(в (сущ., мн.ч., тв.п.)	null	null	null	null
13	(в (сущ., мн.ч., п.п.)	null	null	null	null	(в (сущ., мн.ч., п.п.)	null	null	(в (сущ., мн.ч., п.п.)	null	null	null	null

Приложение 2. МаПС для примера 3.

Таблица.2-1. МаПС для примера 3.

	что	**вам**	**нравиться**	**в**	**ваших**	**телефонах**
	1	2	3	4	5	6
1	(что (личн. мест., ед.ч., им.п.))	(вам (личн.мест., мн.ч., д.п.)	(нравится (глаг., ед.ч., 3 л., наст.вр., изъяв.))	(в (предл.))	(вашу (прит. мест., ед.ч., ж.р., в.п.))	(телефонах (сущ., м.р., мн.ч., п.п.))
2	(что (личн. мест., ед.ч., в.п.))	null	нравиться (глаг., инф.)	(в (сущ., ед.ч., им.п.)	(ваших (прит. мест., мн.ч., р.п.))	null
3	(что (нар.))	null	(нравься (глаг., 2 л., ед.ч., импер.))	(в (сущ., ед.ч., р.п.)	(ваших (прит.мес т., мн.ч., в.п.))	null
4	(что (част.))	null	(нравишьс я (глаг., ед.ч., 2 л., наст.вр., изъяв.))	(в (сущ., ед.ч., д.п.)	(ваших (прит.мес т., мн.ч., п.п.))	null
5	null	null	(нравился (глаг., ед.ч., 3 л., прош.вр., м.р., изъяв.))	(в (сущ., ед.ч., в.п.)	(ваше (прит.мес т., ед.ч., ср.р., им.п.))	null
6	null	null	(нравимся (глаг., мн.ч., 2 л., наст.вр., изъяв.))	(в (сущ., ед.ч., тв.п.)	(ваше (прит.мес т., ед.ч., ср.р., в.п.))	null
7	null	null	null	(в (сущ., ед.ч., п.п.)	(ваших (сущ., ж.р., мн.ч., п.п.))	null
8	null	null	null	(в (сущ., мн.ч., им.п.)	(ваши (прит.мес т., мн.ч., им.п.))	null
9	null	null	null	(в (сущ., мн.ч., р.п.)	(ваши (прит.мес т., мн.ч., в.п.))	null
10	null	null	null	(в (сущ., мн.ч., д.п.)	(вах (межд.))	null

	что	вам	нравиться	в	ваших	телефонах
	1	2	3	4	5	6
11	null	null	null	(в (сущ., мн.ч., в.п.)	(вахш[8] (сущ., ед.ч., м.р., им.п.))	null
12	null	null	null	(в (сущ., мн.ч., тв.п.)	(вахш (сущ., ед.ч., м.р., в.п.))	null
13	null	null	null	(в (сущ., мн.ч., п.п.)	(ваш (прит.мес т., ед.ч., м.р., им.п.))	null
14	null	null	null	null	(ваш (прит.мес т., ед.ч., м.р., в.п.))	null
15	null	null	null	null	(ваша (прит.мес т., ед.ч., ж.р., им.п.))	null

Таблица 2-2.

Словоформы, удаленные на различных этапах редукции матрицы МаПС

Маркировка ячейки	Этап, на котором редуцируется словоформа
	Зачеркиванием с левого верхнего угла в правый нижний выделены ячейки, содержащие словоформы, удаленные на этапе «Очистка словаря».
	Зачеркиванием с левого нижнего угла в правый верхний выделены ячейки, содержащие словоформы, удаленные на этапе «Фильтрация с помощью словаря биграмм».
	Зачеркиванием крест на крест выделены словоформы, удаленные на этапе «Построение грамматического вектора».

[8] Вахш - река в Таджикистане

Приложение 3. Потенциальные ребра, отобранные в результате чанкинга (для примера 3).

1) {chunkId=1, template={id=231}, mainWord={text='Что',correction='**что**',wordform=[pos: "NPRO" lemma: "что" grammems: ["neut", "sing", "nomn"]],wordInd=0}, dependentWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, isHeadChunk=true}

2) {chunkId=2, template={id=233}, mainWord={text='Что',correction='**что**',wordform=[pos: "NPRO" lemma: "что" grammems: ["neut", "sing", "nomn"]],wordInd=0}, dependentWord={text='вашх',correction='**ваша**',wordform=[pos: "ADJF" lemma: "ваш" grammems: ["femn", "sing", "nomn", "Apro"]],wordInd=4}, isHeadChunk=true}

3) {chunkId=3, template={id=37}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

4) {chunkId=4, template={id=43}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

5) {chunkId=5, template={id=44}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

6) {chunkId=6, template={id=48}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

7) {chunkId=7, template={id=50}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='**телефонах**',correction='телефонах',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

8) {chunkId=8, template={id=55}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

9) {chunkId=9, template={id=56}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

10) {chunkId=10, template={id=57}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

11) {chunkId=11, template={id=58}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

12) {chunkId=12, template={id=65}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

13) {chunkId=13, template={id=68}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

14) {chunkId=14, template={id=69}, mainWord={text='нравиься',correction='**нравится**',wordform=[pos: "VERB" lemma: "нравлюсь" grammems: ["sing", "impf", "intr", "3per", "pres", "indc"]],wordInd=2}, dependentWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, isHeadChunk=true}

15) {chunkId=15, template={id=1}, mainWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, dependentWord={text='вашх',correction='**ваших**',wordform=[pos: "ADJF" lemma: "ваш" grammems: ["plur", "loct", "Apro"]],wordInd=4}, isHeadChunk=false}

16) {chunkId=16, template={id=4}, mainWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, dependentWord={text='вашх',correction='**ваших**',wordform=[pos: "ADJF" lemma: "ваш" grammems: ["plur", "loct", "Apro"]],wordInd=4}, isHeadChunk=false}

17) {chunkId=17, template={id=5}, mainWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, dependentWord={text='вашх',correction='**ваших**',wordform=[pos: "ADJF" lemma: "ваш" grammems: ["plur", "loct", "Apro"]],wordInd=4}, isHeadChunk=false}

18) {chunkId=18, template={id=6}, mainWord={text='телефонах',correction='**телефонах**',wordform=[pos: "NOUN" lemma: "телефон" grammems: ["inan", "masc", "plur", "loct"]],wordInd=5}, dependentWord={text='вашх',correction='**ваших**',wordform=[pos: "ADJF" lemma: "ваш" grammems: ["plur", "loct", "Apro"]],wordInd=4}, isHeadChunk=false}

Информация об авторе:

Поляков Владимир Николаевич

к.т.н., доцент, НИТУ «МИСиС», Москва, Ленинский пр-т, 4

с.н.с., Институт Языкознания РАН, Москва, Большой Кисловский п., 1

pvn-65@mail.ru