Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей

Кудинов М. С.
mikhailkudinov@gmail.com
Москва,
Вычислительный Центр им. А. А. Дородницына РАН.
Управление высокопроизводительных алгоритмов Исследовательского центра «Самсунг»

В статье изложен подход к поиску синтаксически связанных групп соседних слов (chunks) в русском тексте. Продемонстрирована принципиальная возможность и корректность постановки задачи выделения таких групп применительно к языку со свободным порядком слов. С использованием аппарата условных случайных полей определенный класс подобных групп можно выделить с F_1 мерой не менее 0.94. При этом обучающая выборка может быть получена путем обработки исходного текста синтаксическим анализатором без последующей ручной коррекции результатов. Тем не менее выделение достаточно длинных фрагментов текста оказывается затруднительным, а показатель F_1 меры, полученный в эксперименте, достаточно низким.

Ключевые слова: *графические вероятностные модели, поверхностный синтаксический разбор, обработка естественного языка.*

Shallow Parsing of Russian Text with Conditional Random Fields

Kudinov M. S.
Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS

The paper describes an aproach to chunking of sentences in Russian. Arguments in favor of the correctness and practicability of the chunking problem for a language with free word order are provided. An aproach based on conditional random fields provides detection of a certain class of chunks (base-NPs) with F_1 measure above 0.94 and the training set can be obtained from the raw text data processed by statistical parser without manual postprocessing. Meanwhile, detecting of longer phrases remains problematic and the F_1 measure in the corresponding experiment is relatively small.

Keywords: *probabilistic graphical models, shallow parsing, natural language processing.*

Введение

Решение задачи синтаксического разбора является одним их ключевых промежуточных пунктов в большом количестве задач, связанных с обработкой естественного языка. Важнейшими из них являются моделирование понимания речи и извлечение фактов из текста. Полный синтаксический разбор основан на аппарате синтаксических деревьев [1]. Однако полный синтаксический разбор имеет ряд недостатков, вследствие которых он не является одинаково пригодным для различных задач обработки естественного языка. К таким недостаткам можно отнести большой объем статистической модели (в случае, если анализатор основан на машинном обучении), большой расход памяти или низкое быстродействие, относительно невысокая точность разбора большинства современных решений. Для организации систем речевого диалога полный синтаксический разбор тем бо-

лее затруднителен, что в устной речи говорящий часто склонен организовывать свою речь не в форме предложений, а в форме более коротких фраз. Кроме того, полный синтаксический разбор является неустойчивым к незнакомым и ошибочно распознанным словам.

В работе [2] был предложен подход, основанный на выделении синтаксически связанных фрагментов текста (прямое заимствование «чанк» — chunk — пока не является распространенным в научной литературе, поэтому далее позволим себе использовать сокращение СФТ.) с помощью каскада конечных автоматов [3] с их последующим объединением в дерево на втором этапе работы алгоритма. Единственным формальным требованием к связанным фрагментам была нерекурсивность. Выделение связанных фрагментов именных групп без построения синтаксического дерева было обозначено в качестве самостоятельной задачи в работе [4]. В этой же работе впервые был предложен подход, основанный на машинном обучении. В работе [5] задача выделения СФТ была сведена к задаче маркировки последовательности слов с использованием СММ.

В работе [6] на примере решения задачи расстановки частеречных тегов, также являющейся задачей маркировки последовательностей, была продемонстрирована высокая эффективность марковской модели максимальной энтропии. В силу того, что ММЭ является дискриминационной моделью и позволяет использовать большое количество, вообще говоря, коррелированных признаков, ее использование значительно улучшило результаты распознавания, по сравнению с СММ.

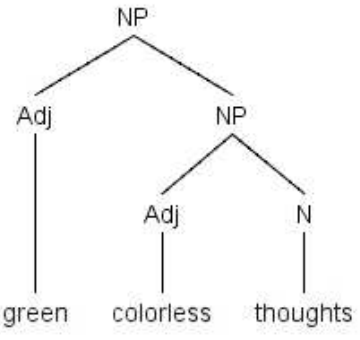
В работе [7] было предложено очередное улучшение модели маркировки последовательностей - условное случайное поле для линейной цепи (Linear-chain CRF, см. [8]). Авторы также использовали аппарат условных случайных полей для задачи расстановки частеречных тегов. Важнейшим преимуществом условных случайных полей перед марковской моделью максимальной энтропии является преодоление т.н. проблемы *смещения метки* (label bias problem), которая заключается в том, что в силу специфики коэффициента нормализации в каждом из факторов цепи для ММЭ состояния, имеющие высокую энтропию распределения последующих состояний имеют меньшие шансы быть выбранными в качестве меток, даже если на это указывают наблюдения [9].

Наконец, в статье [10] аппарат условных случайных полей был успешно применен к выделению СФТ в английском тексте. В качестве целевых меток использовались метки BIO, обозначающие начало фрагмента, середину фрагмента и нахождение вне фрагмента соответственно.

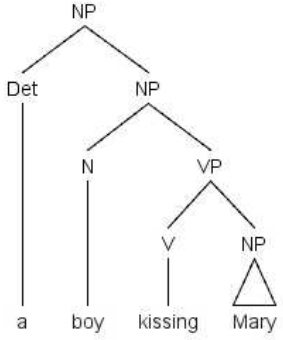
Использование методов машинного обучения в задачах выделения СФТ на материале языков с развитой морфологией (в том числе, славянских) и, как следствие, более свободным порядком слов на настоящий момент не распространено. Обзор методов поверхностного синтаксического разбора для чешского и польского языков изложены в работах [11], [12]. В обоих случаях использован подход, основанный на правилах.

При попытке описать существенный фрагмент грамматики языка разработка правил становится все более трудоемкой, а описание, иногда противоречивым, что приводит к необходимости ручного ранжирования. Этих проблем удается избежать, если использовать методы машинного обучения. Для начала, однако, необходимо уточнить формулировку задачи с учетом специфики языка.

Таким образом, в разд. 1 будут сформулированы и переформулированы некоторые базовые термины, необходимые для дальнейшего изложения. Далее в разд. 2 будут приведены данные, косвенно свидетельствующие о корректности и целесообразности решения задачи определения СФТ на русском материале в принципе, а также приведенных выше формулировок. В разд. 3 кратко излагается аппарат условных случайных полей. В



(а) Именная группа, являющаяся базовой именной группой



(б) Именная группа, не являющаяся базовой именной группой (нарушено условие нерекурсивности)

Рис. 1: Связь между проективностью и возможностью выделения базовых именных групп: построение базовой именной группы длинной более одного слова для дерева (б) невозможно

разд. 4 содержатся информация об обучающих и тестовых данных. В разд. 5 содержится информация о реализации алгоритмов и результаты экспериментов.

Базовые именные группы в русском языке

Центральным понятием в задаче выделения СФТ является понятие базовой группы, или base-XP. XP является обозначением синтаксической составляющей, принятым в англоязычных статьях. Так синтаксическая составляющая, возглавляемая вершиной-существительным, называется NP — noun phrase (именной группой); для глагола это VP — verb phrase (глагольная группа); синтаксическую составляющую безотносительно к ее типу принято обозначать как XP.

В работе [2], на возможные СФТ накладывалось следующее формальное ограничение: в их полном синтаксическом разборе не должно было присутствовать рекурсивных правил. Было продемонстрировано, что такие фрагменты выделяются значительно проще ввиду меньшего влияния синтаксической неоднозначности (attachment ambiguity). Таким образом, в первом приближении СФТ представляет собой вершину в синтаксическом дереве с зависимыми, соседними в линейном порядке, причем в полученном поддереве зависимостей не должно быть иных вершин, имеющих ту же часть речи, что и корень. Например:

green colorless thoughts является СФТ

a boy kissing Mary не является СФТ

В работе [4] речь идет об СФТ с вершиной-существительным. Для таких фрагментов было предложен термин базовой именной группы (Base-NP). В [1] отмечается, что строгого определения ни для СФТ (chunk), ни для базовой именной группы (Base-NP) не было предложено, однако требование нерекурсивности оставалось.

Поскольку в дальнейшем речь пойдет именно о базовой именной группе, сформулируем ее определение более строго:

Определение 1. *Последовательность слов в предложении S называется базовой именной группой, если (1) она является подпоследовательностью некоторой именной группы в дереве непосредственных составляющих; (2) в выводе данной подпоследовательности*

в грамматике языка G нетерминальный символ NP, соответствующий вершине именной группы, встречается единственный раз.

Корректность данного определения следует из определения именной группы (см. [13]).

Для того, чтобы дать определение базовой именной группы в русском языке, необходимо учесть тот факт, что английские определения имеют тенденцию помещаться слева в линейном порядке от вершины в синтаксическом дереве. В этом случае принято трактовать существительные как прилагательные:

USA Supreme Court

В русском подобные конструкции часто переводятся с помощью родительного падежа:

Верховный суд Российской федерации

Это подсказывает определение базовой именной группы в русском языке:

Определение 2. *Последовательность слов в предложении S на русском языке называется базовой именной группой, если 1) Она является подпоследовательностью некоторой именной группы в дереве непосредственных составляющих; 2) В дереве непосредственных составляющих для данной подпоследовательности существует не более одной именной группы, в т. ч. вершина данной группы стоит не в родительном падеже.*

Данное определение, безусловно, не покрывает всех возможных случаев употребления родительного падежа, но учитывает два наиболее важных случая употребления: 1) обозначение синтаксической зависимости между двумя существительными и 2) употребление родительного падежа в отрицательных предложениях.

Ниже будет показано, что выделение базовых именных групп согласно определению 2, действительно, позволяет выделять в тексте логически связанные фрагменты, однако вначале приведем экспериментальные свидетельства того, что более свободный порядок слов в русском языке не является «фатальным» для выделения связанных фрагментов текста.

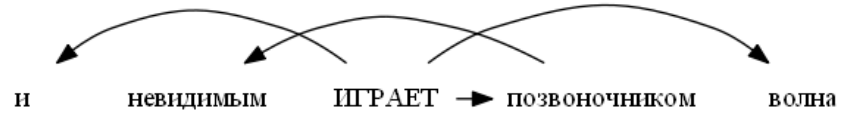
Эффекты свободного порядка слов

Одной из наиболее трудных проблем, которые ставит перед инженерами по обработке ЕЯ русский язык, является свободный порядок слов. Тем не менее имеется ряд как теоретических аргументов [13], так и эмпирических фактов, свидетельствующих об обратном. Так, известно, что большинство предложений на русском литературном языке проективны. Это, в свою очередь, означает, что в русских предложениях должна прослеживаться тенденция к сохранению базовых именных групп, хотя проективность сама по себе не гарантирует сохранение синтаксических групп. Для оценки эффектов от свободного порядка слов был поставлен эксперимент на небольшом корпусе интервью в электронных СМИ.

Обрабатываемые данные Обрабатываемый корпус составляли 500 предложений, взятых из интервью, опубликованных в интернет-изданиях. Каждый текст предварительно обрабатывался последовательно морфологическим анализатором TreeTagger [14], лемматизатором CSTlemma [15] и синтаксическим анализатором Malt Parser [16]. В результате был получен синтаксически аннотированный корпус, где каждому предложению было сопоставлено дерево зависимостей.

Экспериментальная методика Каждое синтаксическое дерево обрабатывалось алгоритмом, который восстанавливал порядок слов, соответствующий сильно проективной конструкции. Все предложения, подвергшиеся изменению, просматривались вручную для выяснения причин изменения исходного порядка слов. Просмотр выявил, что большая часть изменений (17 случаев) была вызвана ошибками в синтаксическом разборе Malt

(а) Дерево зависимостей и границы базовых именных групп для проективного предложения



(б) Дерево зависимостей для непроективного предложения

Рис. 2: Связь между проективностью и возможностью выделения базовых именных групп: построение базовой именной группы длинной более одного слова для дерева (б) невозможно

Parser'а. Из оставшихся 9 случаев 6 приходились на слабо-проективные конструкции с составным глагольным сказуемым (например, *захотел пойти*) и не могли оказать влияния на базовые именные группы. В результате лишь 3 случая приходились на действительно непроективные конструкции. Однако даже среди них дважды встретилась конструкция *друг к другу*, которая была разобрана как полноценное синтаксическое поддерево, что вообще говоря спорно.

Безусловно, экспериментальная выборка не может считаться репрезентативной, однако она позволяет предположить, что свободный порядок слов не является причиной для отказа от выделения СФТ именных групп. Одним из подходов к выделению таких фрагментов, показавший свою эффективность для многих языков со строгим порядоком слов, является подход, основанный на условных случайных полях для линейной цепи.

Условные случайные поля для линейных цепей

Условное случайное поле представляет собой частично ориентированную графическую вероятностную модель, а именно марковскую сеть, в которой при этом имеется условное распределение одного подмножества переменных (ненаблюдаемых) в зависимости от другого подмножества переменных (наблюдаемых). Условные случайные поля широко используются в сегментации изображений, распознавании действий, обработке естественного языка и других областях [9]. Общее определение для условных случайных полей таково:

Определение 3. *Условным случайным полем для переменных* $\boldsymbol{X} \cup \boldsymbol{Y}$ *называется неориентированный граф* H*, с множеством вершин* $V = \boldsymbol{X} \cup \boldsymbol{Y}$ *и множеством ребер с ассоциированными факторами* $\varphi_1(\boldsymbol{D}_1), \ldots, \varphi_m(\boldsymbol{D}_m)$ $(\varphi_i(\boldsymbol{D}_i) \geqslant 0)$*, причем* $\forall i \boldsymbol{D}_i \nsubseteq \boldsymbol{X}$*, а распределение вероятностей* $\mathsf{P}(\boldsymbol{Y} \mid \boldsymbol{X})$ *задается согласно формулам:*

$$\begin{aligned} \mathsf{P}(\boldsymbol{Y} \mid \boldsymbol{X}) &= \frac{1}{Z(\boldsymbol{X})} \tilde{\mathsf{P}}(\boldsymbol{Y}, \boldsymbol{X}) \\ \tilde{\mathsf{P}}(\boldsymbol{Y}, \boldsymbol{X}) &= \prod_{i=1}^{m} \varphi_i(\boldsymbol{D}_i) \\ Z(\boldsymbol{X}) &= \sum_{\boldsymbol{Y}} \tilde{\mathsf{P}}(\boldsymbol{Y}, \boldsymbol{X}) \end{aligned} \tag{1}$$

Тогда любые две вершины y_k, y_l *в* H *соединены неориентированным ребром тогда и только тогда, когда* $\exists \varphi_i \{y_k, y_l\} \subseteq \boldsymbol{D}_i$

В настоящей статье нас интересуют условные случайные поля определенного типа, а именно условные случайные поля для линейной цепи. Данную модель можно рассматривать как дискриминационный аналог скрытой марковской модели. Поэтому будем считать, что $\mathbf{x}$ есть последовательность слов в предложении, а $\mathbf{y}$ - последовательность меток BIO. Соответствующее условное случайное поле выглядит так, как показано на рис. 3.

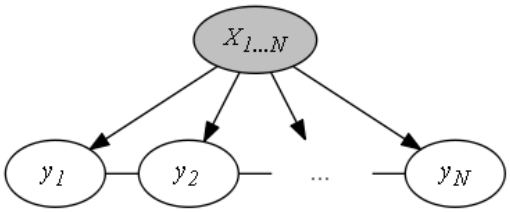


Рис. 3: Условное случайное поле для линейной цепи. Цветом обозначаются наблюдаемые значения. Дуги соответствуют факторам. Прописное X и направленные дуги указывают на то, что значение каждого фактора $\varphi_t^2(y_t, y_{t-1})$ пропорционально вероятности $p(y_t, y_{t-1}|X_{1...N})$, т.е. вероятности того, что данные скрытые состояния «участвовали» в генерации всей наблюдаемой последовательности. Подробнее о нотации см. [9]

В данной цепи имеются два типа факторов - одиночные (singleton) и парные (pairwise). Одиночные факторы $\varphi_t^1(y_t, \mathbf{x})$ задают влияние, которое оказывает наблюдаемая последовательность на метку y_t. Парные факторы $\varphi_t^2(y_t, y_{t-1})$ задают влияние соседних меток друг на друга. Важным отличием от СММ и ММЭ является то, что факторы вообще говоря не обязаны удовлетворять неравенству $\varphi(\mathbf{D}) \leqslant 1$. Аргумент $\mathbf{x}$ в факторе φ_t^1 указывает на то, что при вычислении его значений потенциально могут использоваться признаки любых элементов последовательности. Такими признаками, например, могут быть «слово x_t является существительным» или «слово x_t 'человек', слово x_{t+1} 'собака'». С другой стороны, φ_t^1 является функцией y_t, поскольку значение $\mathbf{x}$ постоянно.

Стандартным методом вывода в графических моделях является конструирование кластерного дерева с последующим применением алгоритма Витерби. В случае линейной цепи кластерное дерево также является цепью. Например, если в структуре на рис. 3 значения в каждом из одиночных факторов будут вычисляться только на основе признаков текущего наблюдения, то кластерное дерево примет следующий вид:

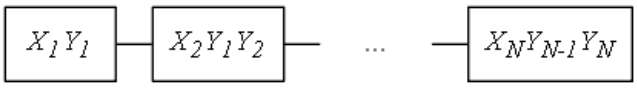


Рис. 4: Кластерное дерево, соответствующее условному случайному полю для линейной цепи с «окном наблюдения» равным 1

Фактор кластера ψ_t представляет собой произведение факторов, входящих в него. Таким образом, его область определения совпадает с областью определения парного фактора, в то время как при вычислении его значений также используются и различные признаки элементов наблюдаемой последовательности. Из этих соображений будем обозначать фактор кластера как $\psi_t(y_t, y_{t-1}, \mathbf{x})$.

Для вычисления значений фактора $\psi_t(y_t, y_{t-1}, \mathbf{x})$ на основе наблюдаемых признаков вводятся индикаторные функции признаков f_k и соответствующие веса λ_k. Тогда искомая

условная вероятность находится по следующей формуле:

$$\mathsf{P}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{X})} \exp\{\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x})\}, \quad (2)$$

где $\mathrm{Z}(\mathbf{X})$ – функция нормализации, которая вычисляется по формуле:

$$\mathsf{P}(\mathbf{y} \mid \mathbf{x}) = \sum_{y} \exp\{\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x})\} \quad (3)$$

На этапе обучения осуществляется подбор весов λ_k, на которых достигается максимум правдоподобия обучающей выборки. Обучение осуществляется с помощью различных вариаций градиентного подъема. Одним из наиболее эффективных является алгоритм L-BFGS. Исчерпывающая информация и ссылки содержатся в [9] и [8].

Используемые данные

Для экспериментов использовались обучающие выборки двух типов. Первая обучающая выборка была получена путем обработки неразмеченного текста из корпуса OpenCorpora морфологическим и синтаксическим анализатором (TreeTagger, CSTlemma, Malt Parser). В результате была получена исходная выборка с синтаксической разметкой в форме дерева зависимостей. Всего 46397 предложений. В качестве второй исходной выборки была использована часть синтаксически аннотированного корпуса русского языка SynTagRus ИППИ РАН. В обучающее множество было выделено всего 40976 предложений. SynTagRus также использует в качестве синтаксической аннотации структуру дерева зависимостей. Далее в каждом синтаксическом дереве выделялись базовые именные группы. Первое слово группы получало метку B, последующие — метку I; слова, не вошедшие ни в одну из базовых именных групп, получали метку O. Алгоритм извлечения базовых именных групп представлял собой простой обход дерева вглубь с объединением в одну базовую именную группу поддеревьев, удовлетворяющих следующим критериям:

1) Все узлы поддерева представляют собой подпоследовательность в линейном порядке слов без разрывов.
2) Вершиной каждого поддерева является слово-существительное в любом падеже.
3) В остальных узлах поддерева могут присутствовать только узлы со следующими характеристиками:
   (а) существительное в родительном падеже;
   (б) прилагательное или порядковое числительное в любом падеже;
   (в) наречие.
4) Знаки препинания отсутствуют в подпоследовательности.

Таким образом, были получены две обучающие выборки с *BIO*-разметкой. В тестовое множество было выделено 6310 предложений корпуса SynTagRus, обработанных тем же способом. Две указанных обучающих выборки и тестовая выборка были использованы в первой серии экспериментов.

Для второй серии экспериментов набор меток *BIO* был расширен двумя метками BH и IH с целью выделения вершины базовой именной группы. Методика подготовки обучающего и тестового множеств осталась прежней, за исключением незначительных необходимых изменений в алгоритме генерации меток на основе синтаксического дерева.

Целью третьей серии экспериментов было оценить возможности алгоритма в выделении более длинных фрагментов текста: рядов однородных членов предложения, конструкций с предлогом или союзом внутри, пунктуации. В этом случае ИГ, вообще говоря, допускает рекурсивную вложенность. Назовем такие фрагменты рекурсивными именными группами. Для этой задачи были проведены эксперименты только с выборкой на основе корпуса SynTagRus.

Эксперименты

Для работы с условными случайными полями была использована библиотека MALLET [17], реализованная на Java. В качестве признаков последовательности были использованы следующие признаки токенов: часть речи, падеж, число, род, заглавная буква. Символы пунктуации рассматривались как самостоятельная часть речи. Для каждого токена в качестве признаков также использовались признаки правого и левого соседа, а также все полученные конъюнкции признаков текущего и каждого из соседних токенов: $\{f_t^1 \& f_{t-1}^1$ $f_t^2 \& f_{t-1}^1 \ldots\}$.

Таким образом, в основном в качестве признаков использовались морфологические характеристики имен. Признак «слово начинается с заглавной буквы» было решено ввести для более качественной обработки имен и должностей: *Президент Российской Федерации Борис Борисович Гребенщиков*. Также было решено поставить отдельный эксперимент для моделей, использующих токен в качестве признака.

Стандартной оценкой качества поверхностного синтаксического разбора является заимствованный из информационного поиска показатель по F_1 мере, вычисляемый на основе доли правильно выделенных СФТ.

Рассмотрим коллекцию текстов, взятых в качестве тестовой выборки. Число СФТ, правильно распознанных алгоритмом, отнесенное к количеству элементов, возвращенных алгоритмом, называется точностью (Precision):

$$Precision = \frac{tp}{tp + fp}, \tag{4}$$

где tp (true positive) — количество СФТ правильно распознанных алгоритмом, fp (false positive) – число СФТ, ошибочно возвращенных алгоритмом.

Полнотой (Recall) называется следующая величина:

$$Recall = \frac{tp}{tp + fn}, \tag{5}$$

где fn (false negative) — число СФТ, ошибочно пропущенных алгоритмом.

Взвешенное среднее гармоническое этих величин называется F_1-мерой:

$$F_1 = \frac{1}{\frac{1}{P} * \frac{1}{2} + \frac{1}{R} * \frac{1}{2}} = \frac{2PR}{P + R} \tag{6}$$

, где P и R – соответственно точность и полнота.

Результаты экспериментов приведены в табл. 1 и 2.

Оценка точности, полноты и F_1-меры для выделения вершин СФТ проводилась аналогичным образом: оценивалась доля вершин, правильно определенных алгоритмом.

Из таблиц видно, что алгоритм демонстрирует сбалансировано высокий результат как по точности, так и по полноте, что соответствующим образом сказывается и на F_1-мере. Более того, расширение набора тегов BIO, используемого, например, для английского

Таблица 1: Результаты экспериментов. Базовые и рекурсивные ИГ.

Модель	Precision	Recall	F_1-мера
`SynTagRus. Базовая ИГ. Токены+`	0.9339	0.9307	0.9323
`SynTagRus. Базовая ИГ. Токены-`	0.9452	0.9427	0.9439
`OpenCorpora. Базовая ИГ. Токены+`	0.9229	0.9115	0.9172
`OpenCorpora. Базовая ИГ. Токены-`	0.9305	0.9238	0.9271
`SynTagRus. Рекурсивная ИГ. Токены+`	0.7521	0.7747	0.7632
`SynTagRus. Рекурсивная ИГ. Токены-`	0.7654	0.7536	0.7594

Таблица 2: Результаты экспериментов. Выделение вершин.

Модель	Precision	Recall	F_1-мера
`SynTagRus. Токены+`	0.9556	0.9514	0.9555
`SynTagRus. Токены-`	0.9648	0.9545	0.9596
`OpenCorpora. Токены+`	0.9601	0.9458	0.9529
`OpenCorpora. Токены-`	0.9523	0.9428	0.9510

языка [10] за счет дополнительных тегов для выделения вершин, нисколько не ухудшает результат, а качество их выделения по F_1-мере также оказывается высоким (см. табл. 2). Это позволяет реализовать возможность постановки всего СФТ именной группы в начальную форму, например, для обращения к базе данных. Также стоит отметить тот факт, что модель обученная на основе выборки, полученной при помощи автоматического синтаксического анализатора, показывает достаточно высокий результат (строки 3, 4 табл. 1). Таким образом, модель может быть обучена на корпусе текстов, собранном самостоятельно, что критично для языков, не имеющих больших открытых лингвистических корпусов (к ним относится и русский). Небольшое снижение всех показателей при добавлении признаков-токенов, по всей видимости, является результатом переобучения. Последним фактом, который нельзя не отметить, является значительное ухудшение как точности, так и полноты при попытке обработки более длинных фрагментов текста. В данном случае гораздо сильнее сказывается синтаксическая неоднозначность. К примеру, для правильного распознавания групп с предлогом внутри использование только морфологических характеристик представляется недостаточным: *[Международный суд по правам человека] в [Гааге] vs. [Международный суд по правам] [человека в Гааге] vs. [Международный суд] по [правам человека] в [Гааге] и т.д.* Вопреки ожиданиям, богатый набор морфологических характеристик словоформ в русском языке не дает желаемого эффекта.

Заключение

В работе была совершена попытка адаптации одной из техник синтаксического анализа, давно и эффективно применяющихся для языков с более фиксированным порядком слов. Несмотря на то, что ранее было принято считать, что свободный порядок слов, является препятствием для применения поверхностного синтаксического анализа, результаты экспериментов показывают, что метод работает достаточно надежно и может найти свое применение в тех задачах, где применение полного синтаксического анализа не требуется. Этими задачами могут быть поиск ключевых слов в документе или высказывании, поступающем на вход диалоговой системы. Метод позволяет выделять в тексте такие фрагменты, как: *[экологическая программа]*, *[Президиум Совета Министров СССР]*, *[бас-*

сейн Азовского моря]. Необходимым условием применения данного метода, является такое специфичное для данного языка определение базовой синтаксической группы, которое бы позволило находить максимально длинные нерекурсивные фрагменты текста, что и было сделано в статье.

Достоинством метода является его нетребовательность к качеству обучающей выборки: она может быть получена на основе выдачи доступного синтаксического анализатора. Безусловным недостатком является низкая точность выделения фрагментов, содержащих внутри предлог и/или союз: *[Комиссия ООН] по [правам человека]* вместо *[Комиссия ООН по правам человека]*. В целом в статье удалось показать, что при внесении необходимых поправок, зависящих от языка, метод работает не хуже, чем для языков с более строгим порядком слов.

Литература

[1] *Jurafsky D., Martin M.* Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. Upper Saddle River, New Jersey: Prentice-Hall, 2009. Pp. 427–458.

[2] *Abney S.* Parsing by chunks. Principle-based Parsing / Eds. R. Berwick, S. Abney, and C. Tenny. Kluwer Academic Publishers, 1991. Pp. 257–279.

[3] *Abney S.* Part-of-speech tagging and partial parsing // Corpus-based methods in language and speech processing. / Eds. S. Young, G. Bloothooft. Kluwer Academic Publishers, 1997. Pp. 124–136.

[4] *Ramshaw L, Marcus M.* Text chunking using transformation-based learning // 3rd Annual Workshop on Very Large Corpora Proceedings, 1995. Pp. 82–94.

[5] *Freitag D., and McCallum A.* Information extraction with HMM structures learned by stochastic optimization // 17th National Conference on Artificial Intelligence (AAAI) Proceedings. Austin, Texas, 2000.

[6] *McCallum A., Freitag M. and Pereira F* Maximum entropy Markov models forinformation extraction and segmentation // 17th International Conference on Machine Learning Proceedings. Stanford, California, 2000. Pp. 591–598.

[7] *Lafferty J., McCallum A. and Pereira F.* Conditional random fields: Probabilistic models for segmenting and labeling sequence data // 18th International Conference on Machine Learning Proceedings. Williamstown, Massachusetts, 2001. Pp. 282–289.

[8] *Sutton C., Mccallum A.* Introduction to conditional random fields for relational learning // Introduction to statistical relational learning / Eds. L. Getoor, B. Taskar. Cambridge, Massachusetts: MIT Press, 2006.

[9] *Koller D., Friedman N.* Probabilistic graphical models: Principles and techniques. Cambridge, Massachusetts: MIT Press, 2009. Pp. 943–961., 345–361., 561–564.

[10] *Sha F, Pereira F.* Shallow parsing with conditional random fields // Proceedings of HLT/NAACL, 2003. Pp. 213–220.

[11] *Dolezalova J, Petkevic V.* Shallow parsing of Czech sentence based on Correct morphological disambiguation // Linguistics Investigations into Formal Description of Slavic Languages / P. Kosta, L. Schürcks L., editors, , 2005.

[12] *Przepiorkowski A.* Slavic information extraction and partial parsing // Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. Praga, 2005.

[13] *Тестелец Я. Г.* Введение в общий синтаксис. Москва: РГГУ, 2001. 800 с.

[14] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/` — TreeTagger — a language independent part-of-speech tagger, 2013.

[15] http://corpus.leeds.ac.uk/mocky/ — Russian statistical taggers and parsers, 2013.

[16] *Sharof S., Nivre, J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. // Proceedings of Dialogue, Russian Conference on Computational Linguistics, 2011.

[17] *McCallum A.* MALLET: A Machine Learning for Language Toolkit. // http://mallet.cs.umass.edu MaltParser, 2002.

[18] *Маннинг К., Рагхаван П., Шютце Х.* Введение в информационный поиск. Москва: Вильямс, 2011. 528 с.

[19] *Грановский Д. В., Бочаров В. В., Бичинева С. В.* Открытый корпус: принципы работы и перспективы // http://opencorpora.org/doc/articles/2010_IMS.pdf