УДК 004.02, 004.822 DOI: 10.26102/2310-6018/2021.32.1.001

Восходящий синтаксический анализ текстов на естественном языке

А.М. Бершадский, П.А. Гудков, Е.М. Подмарькова

Пензенский государственный университет,
Пенза, Российская федерация

Резюме: Актуальность работы обусловлена необходимостью автоматизации процесса принятия решений по юридическим вопросам в различных областях человеческой деятельности. В связи с этим, данная статья направлена на раскрытие подхода к организации процесса синтаксического анализа текстов на естественном языке для последующего автоматического построения семантической сети в соответствии с заданными входными документами. В качестве предметной области рассматривается сфера правовой информации. Предлагаемый авторами подход открывает широкие возможности по смысловому анализу правовых документов и их сравнении между собой. В статье приводится алгоритм восходящего синтаксического разбора. Результаты работы рассмотренного алгоритма применимы для последующего формирования базы знаний по имеющимся текстам правовых документов. В качестве модели представления знаний предполагается использовать семантические сети, что открывает широкие перспективы по автоматизации обработки правовой информации. Помимо решения часто встречающихся на практике задач принятия решений по юридическим вопросам, рассмотренный подход позволит автоматизировать решение такой трудоёмкой задачи, как автоматизация проведения юридической экспертизы нормативно-правовых актов. Проведение этой процедуры необходимо для того, чтобы принимаемые нормативные правовые акты соответствовали принципам допустимости и правомерности включения их в действующую систему права.

Ключевые слова: синтаксический анализ, юридические документы, восходящий разбор, объединение лексем, анализ текста, естественный язык, семантическая сеть, алгоритм.

Для цитирования: Бершадский А.М., Гудков П.А., Подмарькова Е.М. Восходящий синтаксический анализ текстов на естественном языке. *Моделирование, оптимизация и информационные технологии.* 2021;9(1). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=902 DOI: 10.26102/2310-6018/2021.32.1.001

Bottom-up syntax analysis for natural language texts

A.M. Bershadsky, P.A. Gudkov, E.M. Podmarkova
Penza State University, Penza, Russian Federation

Abstract: The need to automate the decision-making process on legal issues in various fields of human activity determines the relevance of this work. In this regard, this article is aimed at disclosing an approach to organizing the process of parsing texts in natural language for the automatic construction of a semantic network corresponding to the given input documents. The subject area is the field of legal information. The approach proposed by the authors opens up wide possibilities for the semantic analysis of legal documents and their comparison with each other. The article discusses the organization of the process of bottom-up parsing natural language texts for the further automatic building a semantic network. The authors propose the text parsing algorithm. Its results are applicable for the further formation of the knowledge base on the available texts of legal documents. Semantic networks are supposed to be used as a model for representing knowledge, which opens up broad prospects for the

automation of legal information processing. In addition to solving the problems of making decisions on legal issues that are often encountered in practice, the considered approach will automate the solution of such a time-consuming task as the automation of the legal examination of regulatory legal acts. The implementation of this procedure is necessary in order for the adopted regulatory legal acts to comply with the principles of admissibility and legality of their inclusion in the current system of law.

Keywords: syntax analysis, legal papers, bottom-up parsing, token concatenation, text analysis, natural language, semantic network, algorithm.

For citation: Bershadsky A.M., Gudkov P.A., Podmarkova E.M. Bottom-up syntax analysis for natural language texts. *Modeling, optimization, and information technology*. 2021;9(1). Available from: https://moitvivt.ru/ru/journal/pdf?id=902 DOI: 10.26102/2310-6018/2021.32.1.001 (In Russ). (In Russ).

Введение

Методы и алгоритмы компьютерной лингвистики, используемые для обработки текстовой информации, применяются во многих сферах человеческой деятельности [1, 2], в том числе и правовой [3].

Центральным понятием системы законодательства любой страны является нормативный правовой акт. Для того, чтобы его подготовить и принять, он должен в обязательном порядке проходить юридическую экспертизу, цель которой – исключить разного рода ошибки и неточности. Проведение экспертизы – занятие довольно сложное, ответственное и трудоемкое, и состоит оно из двух этапов. Первый этап представляет собой исследовательскую деятельность, направленную на всестороннее изучение объекта. Второй этап подразумевает написание заключительного документа. Проведение этой процедуры необходимо для того, чтобы принимаемые нормативные правовые акты соответствовали принципам допустимости и правомерности включения их в действующую систему права [4].

Аналитическая работа, проводимая в рамках первого этапа характеризуется определенным набором методов, таких как формально-юридический, системный, логический, сравнительно-правовой и др. В большинстве своем представляет собой рутинную работу с текстовой информацией. Для автоматизации процесса обработки текста на естественном языке существует множество методов и алгоритмов. Рассматривая процесс анализа текстовых данных в целом, следует отметить его высокую сложность [5]. Любая система анализа текста должна выполнять анализ с точки зрения синтаксиса (структуры предложений), семантики (понятий, применяемых в тексте) и прагматики (правильности употребления понятий и целей их употребления). Текстовый документ при этом последовательно проходит следующие этапы обработки [6]:

- Графематический анализ – для выделения в тексте отдельных структурных единиц: основного текста, заголовков, абзацев, предложений, отдельных слов и др. В ряде случаев здесь же проводится и предморфологический анализ – объединение неразрывных неизменяемых словосочетаний в одну единицу: "что-то", "таким образом" и т.д.
- Морфологический анализ – для определения нормальной формы, от которой была образована данная словоформа, и набора параметров, её характеризующих (часть речи, число, род, падеж).
- Синтаксический анализ – наиболее сложная часть анализа, при котором строится дерево разбора предложения, показывающее взаимосвязи между отдельными словами.

- Семантический анализ – смысловой анализ текста с учетом значений слов, при котором уточняются связи, которые не были выявлены на этапе синтаксического анализа.

После выполнения перечисленных этапов анализа формируется описание на некотором языке внутреннего представления системы [7, 8]. Рассматриваемый в данной статье процесс синтаксического анализа позволяет автоматически формировать семантические сети в соответствии с заданными входными документами. Это открывает широкие возможности по смысловому анализу правовых документов и их сравнении между собой.

В силу сложности и неоднозначности естественного языка нисходящие методы синтаксического разбора текстов представляются неперспективными. В данной статье авторы рассматривают подход восходящего синтаксического анализа, при котором не требуются заранее описанные грамматики естественного языка (которых в силу сложности естественного языка и не существует). Вместо этого используются эвристические правила, которые позволяют группировать отдельные взаимосвязанные языковые конструкции, тем самым формируя иерархическое дерево разбора отдельно взятых предложений текста.

Предлагаемый алгоритм

При рассмотрении данного алгоритма мы предполагаем, что входной текст уже прошёл этапы лемматизации и морфологического разбора. Соответственно, те предложения текста, которые поступают на вход синтаксического анализатора, представляют собой последовательность лексем – отдельных слов и знаков препинания. Для каждого из предложений выполняются следующие шаги алгоритма:

1. Объединение лексем, выполняющееся безошибочно. То есть объединение отдельных слов, которые не допускают двоякого толкования. При этом проверяется, что объединяемые лексемы находятся в одном падеже. На этом шаге обрабатываются прилагательные и местоимения:
   - Объединение прилагательных в скобках, поясняющих основное прилагательное, в одну лексему. Часто встречающийся пример – "законодательные (представительные) органы субъектов Российской Федерации".
   - Объединение нескольких идущих друг за другом прилагательных. Например, демократическое федеративное правовое государство. Слова "демократическое федеративное правовое" при этом будут также объединены в одну группу.
   - Объединение вместе существительного и стоящего перед ним прилагательного, если их падеж совпадает. В результате выполнения этого и предыдущего пунктов такие выражения как "демократическое федеративное правовое государство" будут далее рассматриваться как одна лексема.
   - Объединение идущих друг за другом местоимения и существительного при совпадении их падежей. Например, в предложении "Носителем суверенитета в Российской Федерации является её многонациональный народ" словосочетание "её многонациональный народ" образует одну лексему.

2. Объединение групп существительных. Под существительными в данном случае мы понимаем не только отдельные существительные, но и объединённые группы лексем с предыдущего шага алгоритма. Например, лексема "её многонациональный народ" будет эквивалентна существительному "народ".

Данный шаг алгоритма отделен от предыдущего схожего шага по тем соображениям, что он не является настолько однозначным. Возможны различные варианты объединений. Например, в предложении "Народ осуществляет свою власть через органы государственной власти и органы местного самоуправления" сначала необходимо объединить в группы слова "органы государственной власти" и "органы местного самоуправления", и только после этого объединить их в один список, который будет являться лексемой для дальнейшего синтаксического разбора.

В других случаях необходима другая последовательность объединения. Например, "защита прав и свобод человека и гражданина". В этом случае сначала выполняется формирование лексем "прав и свобод" и "человека и гражданина", и лишь затем эти две лексемы объединяются в одно словосочетание.

Итак, формирование групп существительных выполняется двумя способами:

- Объединение существительных, находящихся в одном падеже и разделенных либо запятой, либо союзами "и" или "или", в одну лексему. Данное объединение выполняется циклически, позволяя формировать списки из более чем двух лексем.
- Объединение двух идущих подряд существительных. При этом возможны следующие варианты:
  - Объединение двух идущих подряд существительных, первое из которых находится в творительном падеже. Например, если рассмотреть фразу "в предусмотренном законом порядке", то при данной операции слова "законом порядке" сократятся до одной лексемы.
  - Объединение двух идущих подряд существительных, второе из которых находится в винительном падеже. Например, "органы власти". При этом, как отмечалось ранее, под существительными понимаются не только отдельные слова, но и лексемы, сформированные на предыдущих шагах алгоритма. Например, словосочетание "органы государственной власти" также будет сведено в одну лексему.
  - Объединение двух идущих подряд существительных, второе из которых находится в родительном падеже. Например, "носителем суверенитета", "власти народа", "присвоение полномочий", "основы конституционного строя" и т.д.
  - Объединение двух идущих подряд существительных, второе из которых находится в творительном падеже. Например, "управлении делами", "распоряжение землей" и т.д.

Поскольку результат синтаксического разбора зависит от порядка применения этих двух правил, то используется следующий подход. Формируется два варианта синтаксического разбора, после чего выбирается наиболее подходящий по критериям наименьшей длины получающегося предложения, количеству уровней вложенности формируемых лексем, числа запятых в списке. Последний критерий является наименее приоритетным.

После того, как идущие подряд существительные объединены, выполняется объединение существительных в скобках. Например, "действиями (или бездействием)", "гражданство иностранного государства (двойное гражданство)", "республика (государство)" и т.д.

3. Перечисленные выше шаги алгоритма не гарантируют того, то свёртка лексем будет выполнена на 100% правильно (в силу неоднозначности естественного языка). Поэтому добавляется дополнительный шаг алгоритма – проверка сформированных лексем на ошибки. В случае их обнаружения выполняется обратная операция – разделение лексем на отдельные слова (за исключением правильного варианта лексемы, в которой была выявлена ошибка), после чего выполняется переход на предыдущий шаг алгоритма.

В настоящий момент реализованы следующие виды проверок на ошибки:

- Выражение в скобках, которое конкретизирует некоторое существительное, оказывается примененным не к одному существительному, а к списку. В этом случае оно применяется к последней лексеме из списка.
- Выражение в скобках, которое конкретизирует некоторое существительное, оказывается примененным не к тому существительному, к которому оно должно относиться. Например, рассмотрим фразу "каждый имеет право пользоваться помощью адвоката (защитника)". Слово "защитник" в скобках должно быть отнесено к существительному "адвокат", а не к лексеме "помощь адвоката". При обнаружении такой ошибки лексема также разгруппируется, а слово в скобках привязывается к нужной части предложения.
- После нескольких итераций группировки лексем в списки может сложиться ситуация, при которой часть элементов списка оказывается несогласованной по падежам с другими лексемами, находящихся на других уровнях иерархии синтаксического дерева разбора. В этом случае происходит разгруппировка несогласованных лексем с сохранением правильных фрагментов.

4. После объединения прилагательных, существительных и местоимений следует этап объединения лексем, связанных по смыслу с помощью тире. Например, в предложении "Российская Федерация – Россия есть демократическое федеративное правовое государство с республиканской формой правления" фраза "Российская Федерация – Россия" будет представлять одну лексему (Рисунок 1).

5. Заключительным этапом предлагаемого алгоритма идёт объединение сформированных лексем с глаголами и краткими причастиями.

В результате выполненных действий получаются простые текстовые конструкции, примеры которых приведены на Рисунках 1 и 2.

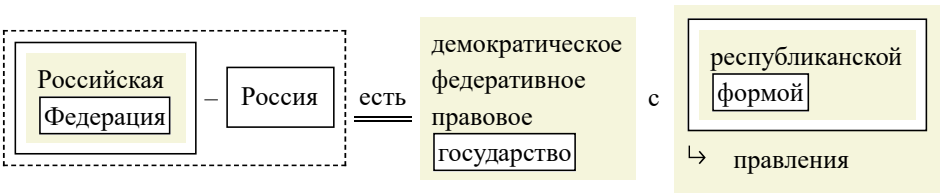

Рисунок 1 – Результат синтаксического разбора предложения «Российская Федерация – Россия есть демократическое федеративное правовое государство с республиканской формой правления»
Figure 1 – Parsing results for the paragraph 1.1 of the Constitution of Russian Federation

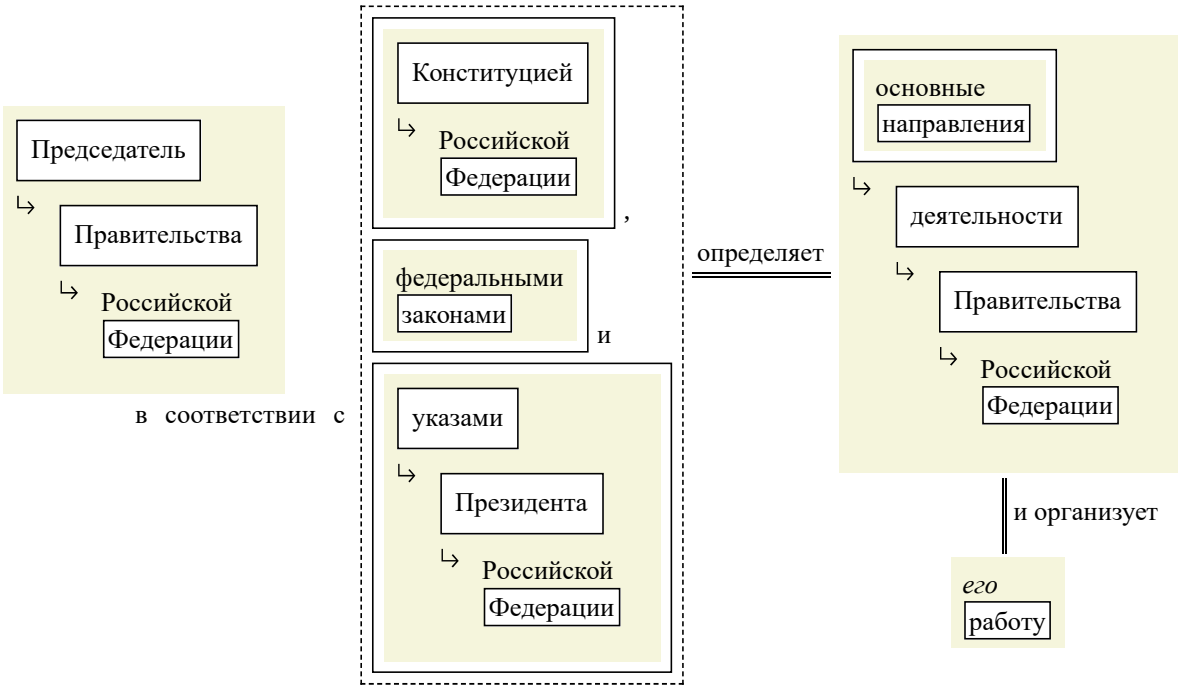

Рисунок 2 – Результат синтаксического разбора предложения «Председатель Правительства Российской Федерации в соответствии с Конституцией Российской Федерации, федеральными законами и указами Президента Российской Федерации определяет основные направления деятельности Правительства Российской Федерации и организует его работу»
Figure 1 – Parsing results for the paragraph 113 of the Constitution of Russian Federation

Получающиеся конструкции, в отличии от исходных предложений, являются более простыми. Для их дальнейшего разбора можно либо использовать методы нисходящего разбора, либо проводить анализ по ключевым словам, либо использовать непосредственно для формирования базы знаний предметной области в виде семантической сети [9].

Заключение

Для того чтобы автоматизировать процесс проведения юридической экспертизы различных документов, требуется выполнять разбор текстов нормативно-правовых актов. Имеются попытки автоматизировать этот процесс, используя простейшие методы анализи текста, такие как поиск по ключевым словам [3]. Авторы же предлагают использовать более глубокий вид анализа, рассматривая входные текстовые документы не только с точки зрения наличия в тексте отдельных конструкций, но и с точки зрения их семантики.

В одной из своих ранних работ [9] авторы рассматривали подход к автоматизированному формированию базы знаний по имеющимся текстам правовых документов. В основе лежал принцип анализа связей на взвешенном графе, описывающем предметную область. Предлагаемый в данной работе алгоритм позволяет

значительно повысить точность распознавания отдельных понятий и их отношений на таком графе за счёт более точного выделения групп взаимосвязанных лексем.

Таким образом, данная работа направлена на решение такой актуальной задачи как автоматизация процесса принятия решений по правовым вопросам в различных сферах человеческой деятельности. Предлагаемый авторами подход, основанный на выполнении восходящего синтаксического разбора естественно-языковых текстов, открывает широкие возможности для семантического анализа юридических документов и их сопоставления между собой.

Тестовая версия разработанной системы, реализующей описанный алгоритм восходящего синтаксического разбора, в настоящее время проходит апробацию на кафедре «Системы автоматизированного проектирования» Пензенского государственного университета.

ЛИТЕРАТУРА

1. Боярский К.К. *Введение в компьютерную лингвистику: учебное пособие.* – СПб: НИУ ИТМО, 2013.
2. Барсегян А.А., Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И. *Анализ данных и процессов: учебное пособие.* – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009.
3. Рахвалова Д.О., Курчеева Г.И., Рахвалова М.Н., Бакаев М.А. Выявление коррупционных факторов в нормативных актах методами крауд-интеллекта. *Государство и граждане в электронной среде* (Труды XXII Международной объединенной научной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19-22 июня 2019 г. Сборник научных трудов). – СПб: Университет ИТМО, 2019(3):66-77. DOI: 10.17586/2541-979X-2019-3-66-77.
4. Пирбудагова Д.Ш. *Юридическая экспертиза нормативных правовых актов: учебное пособие* / под ред. Пирбудагова Д.Ш. 2-е изд., перераб. и доп. Махачкала: Изд-во ДГУ, 2017.
5. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. *Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие.* – М.: МИЭМ, 2011.
6. Батура Т.В. *Математическая лингвистика и автоматическая обработка текстов: учебное пособие*. – Новосибирск: РИЦ НГУ, 2016.
7. Russell S., Norvig P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.
8. Feigenbaum E., Avron Barr. *The Handbook of Artificial Intelligence, Volume III*. Addison-Wesley, 1986.
9. Гудков П.А., Подмарькова Е.М. Модель представления знаний в области правовой информации. *Известия высших учебных заведений. Поволжский регион. Технические науки*. 2020;3(55):17-25. DOI: 10.21685/2072-3059-2020-3-2.

REFERENCES

1. Boyarsky, K.K. *Introduction to computational linguistics: textbook manual.* – SPb: NIU ITMO, 2013.
2. Barseghyan A.A., Kupriyanov M.S., Cold I.I., Tess M.D., Elizarov S.I. *Analysis of data and processes: textbook manual*. – 3rd ed., Rev. and add. – SPb .: BHV-Petersburg, 2009.
3. Rakhvalova D.O., Kurcheeva G.I., Rakhvalova M.N., Bakaev M.A. Revealing corruption factors in normative acts by crowd intelligence methods. *State and citizens in the electronic environment* (Proceedings of the XXII International Joint Scientific Conference "The Internet and Modern Society", IMS-2019, St. Petersburg, June 19-22, 2019 Collection of

scientific papers). – SPb: ITMO University, 2019(3):66-77. DOI: 10.17586/2541-979X-2019-3-66-77.
4. Pirbudagova D.Sh. *Legal examination of regulatory legal acts: textbook manual* / ed. Pirbudagova D.Sh. 2nd ed., Rev. and add. Makhachkala: DSU Publishing House, 2017.
5. Bolshakova E.I., Klyshinsky E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. *Automatic processing of texts in natural language and computational linguistics: textbook manual.* – M .: MIEM, 2011.
6. Batura T.V. *Mathematical linguistics and automatic text processing: textbook manual*. – Novosibirsk: RITs NSU, 2016.
7. Russell S., Norvig P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.
8. Feigenbaum E., Avron Barr. *The Handbook of Artificial Intelligence, Volume III*. Addison-Wesley, 1986.
9. Gudkov P.A., Podmar'kova E.M. Model of knowledge representation in the field of legal information. *Izvestia of higher educational institutions. Volga region. Technical science*. 2020;3(55):17-25. DOI: 10.21685/2072-3059-2020-3-2.

ИНОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Бершадский Александр Моисеевич, д.т.н., профессор, кафедра «Системы автоматизированного проектирования», ФГБОУ ВО «Пензенский государственный университет», Пенза, Российская Федерация.
e-mail: bam@pnzgu.ru
ORCID: 0000-0001-9467-4206

Alexander M. Bershadsky, Doctor of Technical Sciences, Professor, CAD Department, Penza State University, Penza, Russian Federation.

Гудков Павел Анатольевич, к.т.н., доцент, кафедра «Системы автоматизированного проектирования», ФГБОУ ВО «Пензенский государственный университет», Пенза, Российская Федерация.
e-mail: p.a.gudkov@gmail.com
ORCID: 0000-0003-1262-2774

Pavel A. Gudkov, Candidate of Technical Sciences, Associate Professor, CAD Department, Penza State University, Penza, Russian Federation.

Подмарькова Екатерина Михайловна, к.т.н., кафедра «Системы автоматизированного проектирования», ФГБОУ ВО «Пензенский государственный университет», Пенза, Российская Федерация.
e-mail: alpha-and-amega@yandex.ru
ORCID: 0000-0001-6274-269X

Ekaterina M. Podmarkova, Candidate of Technical Sciences, CAD Department, Penza State University, Penza, Russian Federation.