DOI: 10.25728/mlsd.2023.1446

МОРФОЛОГИЧЕСКИЙ И ЛЕКСИЧЕСКИЙ УРОВНИ СИСТЕМЫ АНАЛИЗА ТЕКСТА НА ОСНОВЕ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ ИЗ НЕЙРОНОВ С ВРЕМЕННОЙ СУММАЦИЕЙ СИГНАЛОВ

Харламов А.А.[1, 2], Бородин Н.С.[2]
[1]Институт высшей нервной деятельности и нейрофизиологии РАН, Москва, Россия
[2]Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
kharlamov@analyst.ru, borodinik.s@gmail.com

Аннотация. Рассматривается вопрос разработки нижних уровней (морфологического – словаря окончаний, и лексического – словаря корневых основ) многоуровневой системы анализа текстов, имеющей единый механизм анализа текста на всех его уровнях. Механизм основан на использовании искусственных нейронных сетей на основе нейронов с временной суммацией сигналов.

Ключевые слова: анализ текста, морфологический и лексический уровни, искусственные нейронные сети, нейроны с временной суммацией сигналов.

Введение

Выявление морфологической информации в процессе анализа текста является основой всего процесса анализа, так как именно морфологическая информация позволяет в дальнейшем провести синтаксический разбор предложения, который в случае формального анализа текста является исходной точкой для формирования семантического представления текста.

В лингвистике морфологический анализ слов предложения текста базируется на системе словарей корневых основ и окончаний, которые (словари), в совокупности с предварительным частиречным анализом, позволяют выявить морфологическую информацию, которая и является результатом этого уровня анализа.

Эти словари давно и хорошо известны лингвистам. Они зависят от языка. Они работают во всех известных системах анализа текстов. Если бы существовали реализованные системы интегрального анализа текстов всех уровней от морфологического до семантического, не было бы нужды обсуждать этот вопрос еще раз.

В работе рассматривается вопрос разработки нижних уровней (морфорлогического и лексического) многоуровневой системы для анализа текстов, имеющей единый механизм анализа текста на всех его уровнях. Этот механизм основан на использовании искусственных нейронных сетей, базирующихся на нейронах с временной суммацией сигналов [1]. Представлены процедуры обучения этих механизмов с формированием словарей уровнеобразующих элементов этих двух уровней: словаря корневых основ на лексическом уровне и словаря окончаний – на морфемном.

Поскольку в системах, основанных на правилах, которые являются эталонными для оценки эффективности процедуры анализа, анализ морфологического уровня дает не всегда однозначные результаты, в таких случаях требуется использование дополнительного механизма снятия омонимии. В этом случае используется механизм выявления расширенной предикатной структуры предложения [2]. В будущем, когда к механизмам анализа на нижних двух уровнях добавятся еще два уровня анализа – синтаксический и семантический (также реализованные на основе искусственных нейронных сетей из нейронов с временной суммацией сигналов), снятие омонимии будет осуществляться более широким по отношению к нижним уровням анализа контекстом верхних уровней. То есть основанный на правилах анализ расширенной предикатной структуры предложения будет в этом случае не актуален.

1. Язык как иерархия уровнеобразующих единиц языка

Научение процедуре обработки текстовой информации (обучение грамоте) происходит не в один прием [3]. Анализ перемежается синтезом. Человек проходит стадии обучения: овладение словом, овладение группой слов, овладение предложением.

С самого начала он анализирует слова, выявляя их общие части (лексический уровень обработки). Потом анализирует флексии, их роль в общении (грамматический уровень). Затем анализируются (синтаксические) группы слов, и выявляется согласование слов в группах (синтаксис). Наконец, человек научается выявлять группы имен, отношение между которыми характеризует семантику языка

– допустимость сочетания имен, которому соответствует сочетаемость объектов и событий мира, описываемого текстами.

1.1. Структура нижних уровней языка системы

Структура языка включает структуру словаря лексем языка (грамматические категории и словоизменения), синтаксическую структуру языка (синтаксические группы, отношения между синтаксическими группами в предложении) и структуру семантики (отношение между парами корневых основ имен).

Структура языка в процессе научения языку формируется в виде словаря уровнеобразующих элементов языка разных уровней от морфемного до семантического. В данной работе мы рассмотрим только представления двух нижних уровней языка – морфемного и лексического.

Все множество слов языка делится на грамматические категории, в которых лексемы отличаются морфологическими признаками, а словоформы – еще и признаком словоизменения. Эти категории выявляются статистикой употребления слов и отдельных частей слов. Формируемый статистический граф лексемы (см. Рис. 1) содержит наиболее частотное ядро (корень), менее частотные корневые основы, состоящие из префикса, корня и суффиксов, и, наконец, еще менее частотные флексии – окончания.

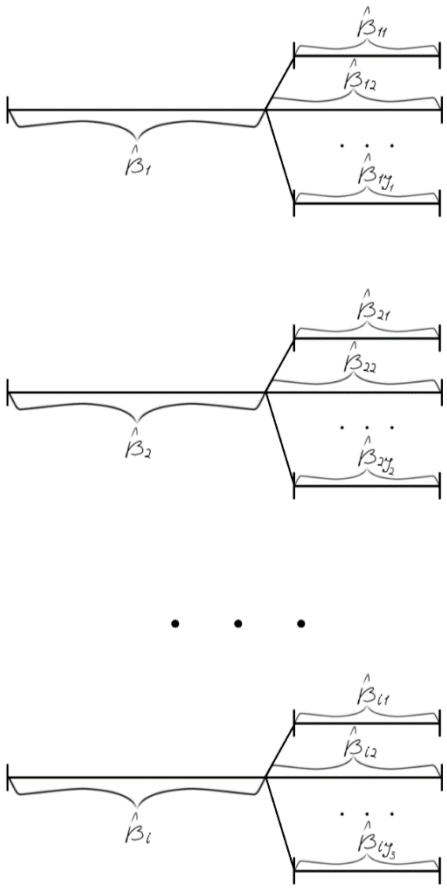

Рис. 1. Графическое представление ядра $\hat{B}_i$ лексемы и окончаний $\hat{B}_{ij}$

Грамматических категорий несколько: существительные, прилагательные, глаголы, наречия, причастия, деепричастия, и подобное. Это классы лексем $\{B_{ij}\}_k$, обладающих основными общими грамматическими свойствами (существительное отличается от глагола по форме).

Словоизменения осуществляются по роду, числу, падежу, склонению, спряжению и еще по другим типам формы, на которые делятся классы грамматических категорий $\{B_{ij}\}$. Они отличаются **словоизменительными окончаниями**.

Таким образом, словарь языка включает в себя несколько классов (и подклассов) слов, каждый из которых представлен корневыми основами, а каждый подкласс представлен различными

словоизменительными окончаниями. **Именно представление этого окончания в заданном подклассе и маркируется морфологическими значениями признаков грамматических категорий**.

При обучении это представление словаря корневых основ для одного класса слов, например существительных $\{B_{iN}\} - \boldsymbol{N - noun}$ формируется, а в процессе использования происходит сравнение входных реализаций словоформ с представлениями подсловарей классов и подклассов. Однозначное или неоднозначное решение позволяет отнести входную словоформу к той или иной грамматической категории.

2. Механизм анализа

2.1. Преобразование в многомерное пространство

Формирование словарей осуществляется [4] ассоциативным преобразованием входного текста в траекторию в многомерном пространстве.

$$\hat{A} = F(A). \tag{1}$$

Здесь F – отображение последовательности A в сигнальное пространство R^n, $\hat{A}$ – траектория:

$$\begin{aligned}\hat{A} = \{\dots, \big(a(-n), a(-n+1), \dots, a(-1)\big), (a(-n+1), a(-n+2), \dots, a(0)), \dots, \\ \big(a(t-n+1), a(t-n+2), \dots, a(t)\big), \dots\} = \\ = (\dots, \hat{a}(-1), \hat{a}(0), \dots, \hat{a}(t), \dots).\end{aligned} \tag{2}$$

2.2. Формирование словаря

Траектория $\hat{A}$ пересекается сама с собой в местах повторов фрагментов входной последовательности *A*. Эти фрагменты траектории $\hat{B}$ являются словарем заданного уровня. То-есть словарь лексем $\{\hat{B}_\text{i}\}$ данного уровня будет включать повторы траекторий данного уровня в количестве, превышающем пороговое значение *h*:

$$\{\hat{B}_\text{i}\} = HM^{-1}MF(\{A\}), \tag{3}$$

где *H* – пороговое преобразование по частоте встречаемости фрагмента траектории, соответствующего слову словаря – повторяющемуся фрагменту входной последовательности. Для этого вводятся *M* – преобразование, соответствующее записи информации в память и M^{-1} – преобразование, соответствующее считыванию из памяти.

2.3. Ассоциативная запись/воспроизведение

Помимо возможности формирования словаря повторяющихся фрагментов входной последовательности, в многомерном пространстве R^n есть возможность осуществлять гетероассоциативную запись произвольной информации по траектории (например, писать код слова или любую произвольную информацию, по траектории слова). Пусть есть две синхронно разворачивающиеся последовательности *A* и *J*. Траектория $\hat{A}$ последовательности *A* (назовем ее несущей) в сигнальном пространстве может быть использована для запоминания в точках пространства R^n , соответствующих траектории, символов синхронизированной с ней информационной последовательности *J*.

Для формирования словаря $\{\hat{B}_\text{i}\}$ введем функцию памяти *M*, ставящую в соответствие каждой точке $\hat{a}(t) \in \hat{A}$, соответствующей *t*-му символу последовательности *A*, двоичную переменную *j(t+1)*, являющуюся *(t+1)*-м символом последовательности *J*.

$$M\{\hat{a}(t), j(t+1)\} = [\hat{a}(t)]_{j(t+1)}. \tag{3}$$

Мы имеем, таким образом, траекторию $\hat{A}$, обусловленную последовательностью *J*. $[*]$ – обозначает обусловленность.

$$[\hat{A}]_J = M\{F(A), J\}. \tag{4}$$

Другими словами, последовательность *J* записывается в точках траектории $\hat{A}$ (в ассоциации с траекторией $\hat{A}$).

Можно осуществить восстановление информационной последовательности *J* по обусловленной ею

траектории $[\hat{A}]_J$ и несущей последовательности A:

$$J = M^{-1}\{[\hat{A}]_J, F(A)\}, \tag{5}$$

где в каждой точке $\hat{a}(t) \in \hat{A}$: $M^{-1}\left([\hat{a}(t)]_{j(t+1)}, a(t)\right) = j(t+1)$. При этом развертывание в траекторию несущей последовательности позволяет обратиться к информации, записанной в точках траектории, то есть к информационной последовательности. Такая запись называется **гетероассоциативной** записью, а воспроизведение – **гетероассоциативным** воспроизведением.

Если в качестве обуславливающей последовательности используется та же последовательность, что и несущая, то есть в точках траектории в сигнальном пространстве записываются символы этой же последовательности, – имеем случай самообусловливания: то есть, если $J \equiv A$, $M^{-1}\{\hat{a}(t), a(t+1)\} = [\hat{a}(t)]_{a(t+1)}$:

$$\left[\hat{A}\right] \equiv [\hat{A}]_A = M\{F(A), A\}. \tag{6}$$

Аналогично (5):

$$A = M^{-1}\{[\hat{A}]_A, F(A)\}. \tag{7}$$

В этом случае можно восстановить последовательность A, начиная с одной из точек ее траектории:

$$A = M^{-1}\{[\hat{A}]_A, \hat{a}(t) \in F(A)\}. \tag{8}$$

Действительно, имея n-членный фрагмент последовательности $\hat{a}(t) = \big(a(t-n+1), a(t-n+2), \dots, a(t)\big)$, мы обращаемся к одной из точек $\hat{a}(t)$ траектории $\hat{A}$. В этой точке записана информация $M^{-1}\{\hat{a}(t), t\} = a(t+1)$, соответствующая следующему символу последовательности A, породившей траекторию $\left[\hat{A}\right]$. Добавляя к *(n-1)*-му символу предыдущего n-членного фрагмента новый символ $a(t+1)$, мы получаем новый n-членный фрагмент $\big(a(t-n+2), a(t-n+3), \dots, a(t+1)\big)$, по которому осуществляется обращение к следующей точке траектории: $\hat{a}(t+2)$. В ней считывается следующий символ последовательности $M^{-1}\{\hat{a}(t+1), t+1\} = a(t+2)$ И так далее до конца последовательности или до ближайшего ветвления траектории. Такая запись называется **автоассоциативной** записью, а воспроизведение - **автоассоциативным** воспроизведением.

2.4. Формирование словарей корневых основ и словоформ

Формирование словаря нижнего уровня (словаря корневых основ) осуществляется частотным анализом повторов элементов этого уровня во входном тексте:

$$\{\hat{B}_i\} = \bigcap_j \{\hat{A}_{i_j}\}, \tag{9}$$

где i – номер лексемы в словаре системы, а $\{\hat{A}_{i_j}\}$ – все анализируемые тексты.

Тогда для одной лексемы B_i граф пересечения траекторий $\hat{B}_{ij}$ всех словоформ этой лексемы будет (ядро лексемы):

$$\hat{B}_i = \bigcap_j \hat{B}_{ij}. \tag{10}$$

Бахрома лексемы (окончания) представляют собой граф дополнения объединения всех словоформ к ядру:

$$\bar{\hat{B}}_{ij} = \bigcup_j \hat{B}_{ij} \setminus \hat{B}_i. \tag{11}$$

3. Алгоритм обработки информации

Формируется словарь лексем как графов с более толстым ядром (корневой основой) и более тонкими ветвями (окончаниями) – см. Рис. 1.

Конкретная словоформа, несущая конкретную морфологическую информацию, имеет также и соответствующую этой морфологической информации флективную структуру (систему окончаний).

Словарь ядер разбивается на классы $\{\hat{B}_i\}_k$, помеченные гетероассоциативно своей

морфологической информацией. *K*- число классов лексем языка.

Для каждого класса $k \in K$ (типа морфологической единицы) формируется своя структура окончаний (в многомерном пространстве) **на следующем уровне** $\{\hat{\bar{B}}_{ij}\}_{k_l}$. Их столько, сколько грамматических классов словоформ (включая их подклассы l внутри класса – типы склонений для существительных, например). Много словоформ – одна флективная структура. Одно ядро вызывает одну или несколько структур.

3.1. Формирование словаря корневых основ в динамическом ассоциативном запоминающем устройстве лексического уровня системы

При обучении формируется словарь ядер словоформ $\{\hat{B}_i\}$, $i \in 1..I$ (в многомерном пространстве) **на нижнем уровне анализа**, где I – число словоформ в классе.

При анализе входное слово (корневая основа+окончание) сразу относится к одному или нескольким классам. Если классов несколько, осуществляется снятие омонимии за счет контекста предложения.

3.2. Формирование словаря флексий в динамическом ассоциативном запоминающем устройстве морфемного уровня

Система словообразовательных окончаний подкласса слов формируется одна для всего подкласса, а встраивается она в модель языка включением индекса веточки словоизменительного графа, соответствующей конкретному окончанию, что соответствует ветвлению выхода нейрона предыдущего уровня на несколько нейронов последующего уровня.

Отдельное окончание представлено в отдельной группе нейронов, и все окончания одного подкласса имеют одинаковый выход – тип подкласса, но со своим словоизменительным вариантом.

4. Реализация

Механизм формирования двухуровневого языкового представления базируется на программной модели искусственной нейронной сети на основе нейронов с временной суммацией сигналов [5].

4.1. Первый уровень обработки на основе динамического ассоциативного запоминающего устройства

Обработка на первом уровне осуществляется при помощи формируемого на основе анализа слов анализируемого текста в динамических ассоциативных запоминающих устройствах [5] (ДАЗУ) этого уровня словаря корневых основ $\{\hat{B}_\text{i}\}$ (Рис. 1). Корневая основа $\hat{B}_\text{i}$ отдельно взятого слова представлена в одном ДАЗУ.

В памяти одного нейрона ДАЗУ хранится комбинация из двух символов: первый символ – буква алфавита или индекс предыдущего нейрона в последовательности нейронов; второй символ – всегда буква алфавита. Таким образом, каждой корневой основе ставится в соответствие уникальная последовательность символов алфавита, хранящихся в нейронах одного из ДАЗУ (фрагмент траектории в многомерном пространстве, моделируемой этим ДАЗУ). В памяти конкретного нейрона также запоминается частота встречаемости хранящейся в нем комбинации как количество обращений к данному нейрону.

Сравнение частоты встречаемости соседних в траектории нейронов позволяет выявить переход от корневой основы к окончанию: разница в частотах говорит о таком переходе.

4.2. Второй уровень обработки на основе динамического ассоциативного запоминающего устройства

На следующем уровне обработки формируется словарь $\hat{B}_{ij}$ следующего уровня: для каждой корневой основы формируется система окончаний (Рис. 1), соответствующая грамматическому классу лексемы.

Для каждого слова словаря $\hat{B}_\text{i}$, относящегося к одному грамматическому классу *k*, система окончаний пишется гетероассоцивтино в его ДАЗУ как еще один элемент памяти, содержащий список окончаний, и для каждого окончания – название его словоизменения (например, падеж – для существительного). Еще в памяти этого нейрона хранится морфологическое название класса *k*. То есть, слово на входе при распознавании приводит к воспроизведению морфологического названия класса *k* и название словоизменительной формы (например, падежа – для существительного).

5. Заключение

В работе рассматривается вопрос разработки нижних уровней (морфорлогического и лексического) многоуровневой системы для анализа текстов, имеющей единый механизм анализа текста на всех его уровнях. Этот механизм основан на использовании искусственных нейронных сетей, базирующихся на нейронах с временной суммацией сигналов. Представлены процедуры обучения этих механизмов с формированием словарей уровнеобразующих элементов этих двух уровней: словаря корневых основ на лексическом уровне и словаря окончаний – на морфемном.

Это представление оказывается неполным, если к нему не добавлено два верхних уровня представления информации: синтаксического и семантического. Создание модели из ДАЗУ, формирующей представление этих уровней – задача последующей разработки, но о них необходимо сказать несколько слов. Синтаксическая обработка, базирующаяся на двух разработанных уровнях, в рамках которых формируется поток морфологической информации, даст возможность выявления расширенной предикатной структуры предложений текста, которая определяет семантику предложений текста и текста в целом, как допустимой попарной сочетаемости корневых основ имен, включенных в предложения текста, соответствующих объектам и событиям мира – субъекта, главного и второстепенных объектов и атрибутов. Именно пары корневых основ имен составляют семантическую сеть текста – модель текста семантического уровня.

Литература

1. Neuroinformatics and Semantic Representations. Theory and Applications. Alexander Kharlamov & Maria Pilgun eds. Cambridge Scholars Publishing. 2020. – 317 p.
2. Осипов Г.С., Смирнов И.В. Семантический анализ научных текстов и их больших массивов. // Системы высокой доступности. № 1. 2016. –С. 41-44.
3. *Лурия А.Р.* Язык и сознание. – С.-Петербург, 2019. – 336 с.
4. Харламов А.А. Ассоциативная память — среда для формирования пространства знаний. От биологии к приложениям. – Дюссельдорф: Palmarium Academic Publishing, 2017. – 109 с.
5. *Alexander Kharlamov* TextAnalyst Technology for Automatic Semantic Analysis of Text. Neuroinformatics and Semantic Representations. Theory and Applications. Collective Monography. Chapter Seven. Cambridge Scholars Publishing. 2020. – P. 182-193.