Нейросетевой подход к распознаванию ситуации по тексту

Харламов А.А.,
доктор технических наук старший научный сотрудник Института высшей нервной деятельности и нейрофизиологии РАН, Москва, профессор Департамента программной инженерии НИУ ВШЭ, Москва, профессор Кафедры прикладной и экспериментальной лингвистики МГЛУ, Москва

Аннотация

Как правило, отдельный текст описывает отдельную ситуацию. Поэтому можно с достаточной степень уверенности соотнести описание ситуации с самой ситуацией. Далее, конкретный текст может быть представлен несколькими способами. Одним из них является формирование семантической сети текста, которая, теперь уже, будет представлять упомянутую ситуацию. Однородная семантическая сеть — это граф, вершины которого представляют собой понятия, а дуги указывают на степень близости этих понятий. Для ее построения используется модель искусственной нейронной сети на основе нейроподобных элементов с временной суммацией сигналов (так называемая кортикоморфная ассоциативная память). Выявление ключевых вершин сети (переранжирование) осуществляется с использованием хопфилдоподобного алгоритма. Используя ассоциативную сеть как представление внутренней структуры текста, мы можем сравнивать тексты по структуре. То есть, что мы можем сравнивать тексты по смыслу, сравнивая между собой их ассоциативные сети. Таким образом, мы можем сравнивать ситуации: чем больше степень пересечения ассоциативных сетей текстов, описывающих ситуации, тем более похожи представленные ими ситуации. Пусть мы имеем множество ситуаций, описанных текстами, которое некоторым образом раскластеризовано на несколько подмножеств. Каждое из этих подмножеств представляет класс ситуаций, чем-то похожих друг на друга. Мы можем решить задачу классификации (распознавания) ситуации, используя алгоритм сравнения сети распознаваемой ситуации с сетями классов.

Ключевые слова: распознавание ситуации, классификация ситуаций, семантическая сеть, сравнение семантических сетей, сравнение смыслов текстов, смысловая классификация текстов, искусственная нейронная сеть, нейроподобный элемент с временной суммацией сигналов, переранжирование понятий семантической сети.

ВВЕДЕНИЕ

Как правило, отдельный текст описывает отдельную ситуацию [1]. Поэтому можно с достаточной степень уверенности соотнести описание ситуации с самой ситуацией. Далее, содержание конкретного текста может быть представлено несколькими способами [2]. Одним из них является формирование семантической сети текста, которая, теперь уже, будет представлять упомянутую ситуацию. А далее, логика проста. Описания ситуаций формируют классы, и текущая ситуация сравнивается с этими классами путем сравнения семантической сети текущей ситуации и семантических сетей классов.

Есть разные подходы к классификации текстов. Почти все они основываются на так называемом тематическом анализа текстов — статистических подходах (LSA, pLSA, LDA), в основе которых лежит монограммная модель текста, а сравниваемые индексы — это перечни тем сравниваемых текстов (то есть их векторные представления) [7]. Более сложным представлением текстов для последующей кластеризации и классификации являются так называемые нейросемантические сети [11], которые включают несколько уровней единообразно представляемой естественно-языковой информации от единиц морфемного уровня, до единиц синтаксического уровня (в отличие от тематического моделирования, где информация представлена уровнем слов). Но и нейросемантические сети являются векторным представлением текстов, хотя и более сложным (многоуровневым). Нейросемантические сети являются развитием тематического моделирования, так как представляют несколько разрозненных тем, но усложненных как в сторону элементов более мелких, чем слова, так и в сторону элементов более сложных (синтаксические конструкции). Но, как и в первом случае (7), эти векторные конструкции, представляющие отдельные элементы текста, не связаны друг с другом.

Развитием представления семантики текста как в сторону использования вместо монограммной модели текста его n-граммной модели, так и в сторону объединения отдельных тематических элементов текста в единое представление, является нейросетевой подход к формированию однородной семантической сети текста, являющейся ее (текста) смысловым портретом [1]. Этот же нейросетевой подход позволяет выявить и тематическую структуру текста, но не как множество отдельных тем, полученных в процессе тематического моделиирования, а как тематическое дерево текста, в котором есть наиболее важная корневая тема, а также темы более низких уровней: подтемы, подподтемы, и подобное. Дополнительным достоинством такого сетевого представления смысла текста является хорошая интерпретируемость семантической сети как множества концептов текста в их взаимосвязях.

1. Однородная семантическая (ассоциативная) сеть текста

Однородная семантическая сеть — это граф, вершины которого представляют собой понятия, а дуги указывают на близость этих понятий в рамках анализируемой структуры. Если однородная семантическая сеть (будем ее называть просто «семантическая сеть») построена на основе анализа некоторого текста, то ее вершины — это понятия текста. Формирование такой семантической сети представлено в другом докладе этого сборника [1]. Для ее построения используется модель искусственной нейронной сети на основе нейроподобных элементов с временной суммацией сигналов (так называемая кортикоморфная ассоциативная память) [2].

Для возможности последующего использования семантической сети в вычислительных алгоритмах оценим смысловые ранги вершин в рамках текста, на основе которого сеть была построена. Выявление ключевых вершин сети (переранжирование) осуществляется с использованием хопфилдоподобного алгоритма [12]. При этом ассоциативная память моделирует обработку в колонках коры больших полушарий, а переранжирование — участие в формировании представления о ситуации, формируемое в ламелях гиппокампа [3]. Достоинством однородной семантической сети является возможность ее автоматического формирования из исходного текста (корпуса текстов). Неоднородные семантические сети формируются вручную [13].

2. Сценарий текста (ситуации)

Ранжирование понятий семантической сети оказывается очень важным ля дальнейшего анализа текста. Подсчет весовых характеристик предложений как нормированных сумм весов включенных в них слов, с последующим применением порогового преобразования, позволяет выявить наиболее важные предложения текста. Они и составляют реферат текста. Реферат далее используется для классификации текстов.

Реферат текста как множество предложений текста в порядке их встречаемости в тексте, ранги которых превышают некоторый заданный порог можно считать сценарием ситуации, представленной в тексте, — предложениями, существенно важными для описания ситуации.

3. Нейросетевой механизм формирования однородной семантической сети

Искусственная нейронная сеть — кортикоморфная ассоциативная память — формирует частотный словарь текста, на основе которого путем сравнения частотных характеристик слов формируемого словаря собираются пары слов, в дальнейшем используемые для построения частотной сети.

Автоматическое формирование частотного словаря анализируемого текста осуществляется программно реализованной иерархической структурой из блоков ассоциативной памяти. Число уровней в иерархической структуре определяет априорно заданную максимально допустимую длину понятия предметной области и равняется двадцати в конкретном случае реализации технологии TextAnalyst [5].

На первом уровне иерархической структуры представлен словарь двухбуквенных слов текста, а также двухбуквенных сочетаний из слов этого словаря. На втором уровне иерархической структуры блоков ассоциативной памяти представляются трехбуквенные слова, а также трехбуквенные сочетания из слов, встреченных в тексте, в виде индексов элементов соответствующих словарей первого уровня, дополненных еще одной буквой. На последующих уровнях представление информации полностью однородно в них хранятся индексы элементов хранения более низкого уровня, дополненные одной буквой.

В процессе формирования представления информации в иерархической структуре блоков ассоциативной памяти подсчитывается частота встречаемости каждого сочетания букв. Частота слов (сочетаний букв, не имеющих продолжения на следующем уровне) используется для последующего анализа.

Сформированное таким образом представление лексики текста подвергается затем пороговому преобразованию по частоте встречаемости. Порог отражает степень детальности описания текста. Таким образом, выявляются устойчивые термины и терминологические словосочетания, которые служат далее в качестве элементов для построения семантической сети.

Семантическая сеть формируется как множество пар слов. В качестве критерия для определения наличия ассоциативной семантической связи между словами пары в анализируемом тексте используется частота их совместной встречаемости в предложениях текста. Непревышение разницей частот следующих друг за другом слов некоторого порога позволяет говорить о наличии между словами ассоциативной (семантической) связи.

Множество пар слов, полученных в результате такого анализа, формирует частотную ассоциативную сеть. Частотная ассоциативная сеть становится семантической сетью после переранжирования частотных характеристик ее вершин с помощью итеративной процедуры.

Далее семантическая сеть используется для сравнения текстов по смыслу и их классификации.

4. Сравнение текстов (ситуаций) по смыслу

Используя ассоциативную сеть как представление структуры текста, мы можем сравнивать тексты по структуре. Будем считать, что структура текста соответствует его смыслу. То есть, что мы можем сравнивать тексты по смыслу, сравнивая между собой их ассоциативные сети [14]. Таким образом, мы можем сравнивать ситуации: чем больше степень пересечения ассоциативных сетей текстов, описывающих ситуации, тем более похожи представленные ими ситуации (см. рис. 1).

Определение 1. Введем скалярное произведение на векторах $\bar{c}_i$, $\bar{c}_j$, где угол между векторами понятий c_i c_j от 0 до 90° пропорционален весу связи от c_i к c_j: w_{ij}. Площадь треугольника s_j, построенного на векторах $\bar{c}_i$, $\bar{c}_j$, развернутых на угол, пропорциональный w_{ij} относительно друг друга, будет использована для вычисления степени пересечения сначала звездочек, а потом семантических сетей как совокупностей звездочек.

Определение 2. Под пересечением двух звездочек z_{i_1} и z_{i_2}, имеющих одинаковое главное понятие c_i, понимается сумма (по всем понятиям-ассоциантам c_j главного понятия c_i этих звездочек) пересечений площадей двух треугольников, построенных в плоскости векторов $\bar{c}_i$, $\bar{c}_j$, один из которых построен на векторах, развернутых на угол, пропорциональный связи w_{ij_1} между понятиями в одной звездочке z_{i_1}, а другой — на угол, пропорциональный связи w_{ij_2} между теми же понятиями в другой, сравниваемой с первой, звездочке z_{i_2}. В случае если в одной из звездочек пары, для которой считается пересечение:

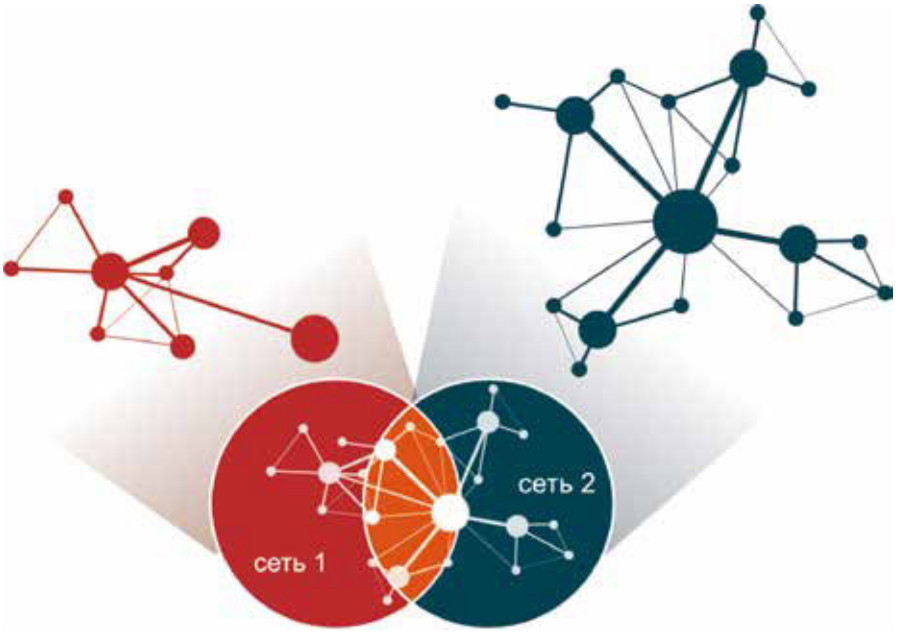

Рис. 1. Степень смыслового подобия текстов (ситуаций) определяется степенью пересечения семантических сетей.

$$s_{12} = s_1 \cap s_2 = < c_{i_1}, c_{j_1} > \cap < c_{i_2}, c_{j_2} > \quad (1)$$
$$= \sum_{j=1}^{\max(n_1 n_2)} min(|\bar{c}_{i_1}||\bar{c}_{j_1}| \cos(\bar{c}_{i_1}, \bar{c}_{j_1}), |\bar{c}_{i_2}||\bar{c}_{j_2}| \cos(\bar{c}_{i_2}, \bar{c}_{j_2})$$

не нашлось соответствующего понятия-ассоцианта, пересечение считается равным 0. Здесь n_1, n_2 — число ассоциантов в звездочках соответственно z_{i_1} и z_{i_2}.

Определение 3. Под пересечением семантических сетей понимается сумма пересечений звездочек, включенных в эти сети (считая по главным понятиям):

$$N_{12} = N_1 \cap N_2 \quad (2)$$
$$= \sum_{i}^{\max(M_1 M_2)} \sum_{j=1}^{\max(n_1 n_2)} min(|\bar{c}_{i_1}||\bar{c}_{j_1}| \cos(\bar{c}_{i_1}, \bar{c}_{j_1}), |\bar{c}_{i_2}||\bar{c}_{j_2}| \cos(\bar{c}_{i_2}, \bar{c}_{j_2})),$$

M_1, M_2 — число звездочек, входящих соответственно в семантические сети N_1 и N_2.

Определение 4. Под классификацией входной ситуации можно понимать отнесение семантической сети ситуации N к сети $\boldsymbol{N_n}$ (где n=1..N — число предметных областей) одной из предметных областей модели мира. В идеальном случае семантическая сеть ситуации вкладывается в сеть соответствующей предметной области.

Используя операцию пересечения сетей N_1 и N_2, определенную выше, мы можем оценивать степень подобия двух сетей $N_1 \cap N_2$ (рис. 2) и, тем

самым, сравнивать по смыслу (по структуре) ситуации (их модели). Имея модели предметных областей в виде ассоциативных семантических сетей, мы можем классифицировать входные ситуации (описывающие их модели) вычислением степени совпадения (вложения) сети входной ситуации и сетей предметных областей (рис. 2), относя входную ситуацию к той предметной области, у которой степень совпадения сети входной ситуации с сетью предметной области окажется выше.

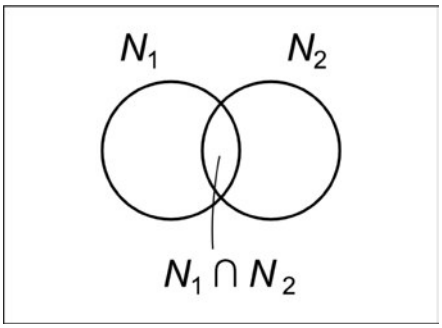

Рис. 2. Пересечение $N_1 \cap N_2$ двух сетей N_1 и N_2, характеризующее степень их смыслового подобия.

5. Классификация текстов (ситуаций)

Пусть мы имеем множество ситуаций, описанных текстами, которое некоторым образом раскластеризовано на несколько подмножеств. Каждое из этих подмножеств представляет класс ситуаций, чем-то похожих друг на друга. Мы можем решить задачу классификации (распознавания) ситуации, используя алгоритм сравнения сети распознаваемой ситуации с сетями классов.

Определение 5. Под классификацией входной ситуации можно понимать отнесение семантической сети ситуации N к сети $\boldsymbol{N_n}$ (где n=1..N — число предметных областей) одной из предметных областей модели мира. Здесь объединение сетей $\cup n\, N_n$ соответствует модели мира. В идеальном случае семантическая сеть ситуации вкладывается в сеть соответствующей предметной области.

Используя операцию пересечения сетей N_1 и N_2, определенную выше, мы можем оценивать степень подобия двух сетей $N_1 \cap N_2$ (рис. 2) и, тем самым, сравнивать по смыслу (по структуре) ситуации (их модели). Имея модели предметных областей в виде ассоциативных семантических сетей, мы можем классифицировать входные ситуации (описывающие их модели) вычислением степени совпадения (вложения) сети входной ситуации и сетей предметных областей (рис. 3), относя входную ситуацию к той предметной области, у которой степень совпадения сети входной ситуации с сетью предметной области окажется выше.

Далее мы можем абстрагироваться от входных ситуаций. Будем рассматривать их языковые модели. Пойдем еще дальше: перейдем к анализу текстов, на основе которых эти языковые модели формируются. Как и в случае классификации входных ситуаций, используя операцию пересечения сетей, мы можем оценивать степень подобия двух сетей и, тем самым, сравнивать по смыслу уже тексты. Имея модели предметных областей N_n в виде ассоциативных семантических сетей соответствующих тематических текстовых выборок, мы можем классифицировать входные тексты вычислением степени совпадения (пересечения/вложения) сети N входного

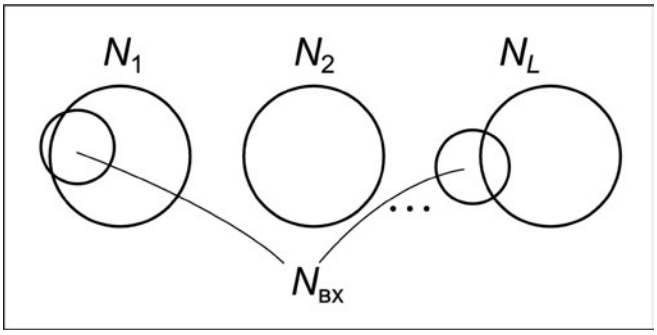

Рис. 3. Классификация входного текста путем выявления степени вложенности его семантической сети N_вх в одну или несколько семантических сетей классов-рубрик — предметных областей N_1, N_2 ... N_l, где l — число предметных областей в модели мира.

текста и сетей предметных областей N_n, относя входной текст к той предметной области, у которой степень совпадения его сети с сетью предметной области окажется выше.

6. Упрощение процедуры классификации

Мы можем упростить процедуру классификации так, как это делают в тематическом моделировании [4]. Выделив минимальный древовидный подграф из ассоциативной сети, мы получим тематическое дерево. Далее мы можем оперировать им так же, как это описано в разделах 2 и 3 для ассоциативных сетей. То есть путем сравнения тематических структур текстов (как в формулах (1), (2), только с заменой сетей на тематические деревья).

7. Усложнение процедуры классификации

Но мы можем и усилить процедуру распознавания, вернувшись вновь к неоднородной семантической сети. Для этого дугам сети назначаются типы отношений между понятиями сети (при этом сеть становится неоднородной как в [4]). Эти типы определяются лингвистическими методами выявлением расширенных предикатных структур соответствующих предложений [5]. Недостатком неоднородной семантической сети является невозможность автоматического выявления всех расширенных предикатных структур текста. Лингвисты в один голос утверждают [9], что автоматически выявляется не более 30 % таких структур. Но даже частично размеченная по типам понятий неоднородная сеть богаче однородной сети.

Заключение

В работе представлен нейросетевой подход к распознаванию ситуации по ее описанию в тексте. При использовании подхода мы заменяем анализ ситуаций анализом их описаний. Представленный подход удобен полностью автоматическим механизмом распознавания (классификации) ситуаций, своей простотой и удобством с точки зрения интерпретации

результатов классификации. Смысловой портрет ситуации (текста) легко интерпретируется как с точки зрения отнесения ситуации к классу (отнесение текста к корпусу текстов, описывающих класс ситуаций), так и с точки зрения выявления общего (для сравниваемых ситуаций) как пересечения семантических сетей.

Литература

1. *Глазков А.В.* Текст как образ мира vs. мир как образ текста // The Peculiarity of Man. — Nr.17 – Toruń-Kielce, 2013. S.97–108.
2. Семантический анализ и способы представления смысла текста в компьютерной лингвистике № 4, 2016. Стр. 45–57.
3. *Ефимова Т.В.* Анализ художественного текста с применением семантической сети 2003
4. *Харламов А.А.* Семантика текста как модель ситуации. В настоящем сборнике.
5. *Харламов А.А.* Ассоциативная память — среда для формирования пространства знаний. От биологии к приложениям. — Дюссельдорф: Palmarium Academic Publishing, 2017. — 109 с. ISBN 978-3-639-64549-1.
6. *Виноградова О. С.* Гиппокамп и память. М.: «Наука», 1975.
7. К*оршунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН (электронный журнал), том 23, 2012. Стр. 215–242.
8. *Alexander A. Kharlamov, Tatyana V. Yermolenko, Andrey A. Zhonin.* Text Understanding as Interpretation of Predicative Structure Strings of Main Text's Sentences as Result of Pragmatic Analysis (Combination of Linguistic and Statistic Approaches)// Доклад на Международной конференции SPECOM 2013, Пльзень, Чехия, сентябрь 2013г. LNAI 8113 — Pp. 333–339.
9. *Городецкий Б.Ю.* Актуальные проблемы прикладной лингвистики // Новое в зарубежной лингвистике. Вып. XII. М., 1983.
10. *Alexander A. Kharlamov, Tatyana V. Yermolenko, and Andrey A. Zhonin.* Modeling of Process Dynamics by Sequence of Homogenous Semantic Networks on the Base of Text Corpus Sequence Analysis// Доклад на Международной конференции SPECOM 2014, Novy Sad, Serbija, September 2014. LNAI 8773 Springer — Pp 300–307.
11. *Харламов А.А., Ле Мань Ха.* Нейросетевые подходы к классификации текстов на основе морфологического анализа // Труды МФТИ. 2017. Т. 9, № 2. С. 143–150.
12. *Hopfield J.J.* Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. 79, 1982. Pp. 2554–2558
13. *Голенков В.В.* Представление и обработка информации в графодинамических ассоциативных машинах. Минск, БГУИР, 2001, 410 с
14. *Харламов А.А., Ермоленко Т.В.* Нейросетевая среда (нейроморфная ассоциативная память) для преодоления информационной сложности. Поиск смысла в слабоструктурированных массивах информации. Часть II. Обработка информации в гиппокампе. Модель мира / Информационные технологии, N 12, 2015. — Стр. 883–889.».

A NEURAL NETWORK APPROACH TO THE SITUATION RECOGNITION BASED ON THE TEXTS

Kharlamov A. A.,
Doctor of Technical Sciences, Senior Researcher, Institute of Higher Nervous Activity RAS, Moscow, Professor, Department of Applied and Experimental Linguistics, MSLU, Moscow, Professor, School of Software Engineering HSE, Moscow

Abstract

Generally, an individual text describes an individual situation. That is why it is possible to correlate a situation with its description with a high degree of certainty. Then, the content of an individual text may be represented in several ways. One of them is the construction of a semantic network of the text. In this case the network will represent the situation. The so called homogeneous semantic network is a graph which nodes represent some concepts and the arcs show the proximity of the concepts within an analyzed structure. If the network is formed as a result of some text analysis then its nodes are the text concepts. For its construction a model of an artificial neural network of neural-like elements with temporal summation (this is so called corticomorphic associative memory) was used as its basis. The nodes ranking is performed by a Hopfield-like algorithm. Thus we can compare the sense of texts by comparing their structure: the larger is the area of the intersection of the associative networks of the texts, the more alike are the represented situations. If there is a set of situations clustered into a set of subsets in some way, then the situations of each subset are included into the situation class with some common features. All the situations of each class resemble each other. Thus, the classification problem may be solved by using the algorithm of comparison between the network of the analyzed situation and those of situation classes.

Keywords: situation recognition, situation classification, semantic network, semantic networks comparison, semantic texts comparison, semantic texts classification, artificial neuronetworks, neuroliked element based on temporal signal summation, semantic network concept rearrangement.