Е. О. Шлычкова, А. Н. Шевляков
Тюменский государственный университет, г. Тюмень
УДК 004.934

АЛГОРИТМ ДЕКОДИРОВАНИЯ СЕМАНТИЧЕСКОЙ ДИСТАНЦИИ НА ОСНОВЕ ДАННЫХ ЭЭГ

Аннотация. Данная работа посвящена определению на биологическом уровне семантической дистанции между словами, высчитанной через расстояние Левенштейна для закодированных с помощью генетического метода дефиниций слов, путем анализа данных электроэнцефалограммы, в том числе с использованием модели машинного обучения (на примере слов испанского языка).

Ключевые слова: ЭЭГ, генетический метод, расстояние Левенштейна, внутренняя речь, семантика.

Введение. Одним из ключевых направлений развития систем декодирования внутренней речи для нейроинтерфейсов «мозг-компьютер», в том числе на основе данных электроэнцефалограмм, является распознавание семантики слов. Данная задача имеет свою специфику ввиду абстрактности значений слов и неопределенности их расположения на уровне головного мозга человека. Мы предполагаем, что подобно семантическим полям для речевых агентов, создаваемых в рамках методов компьютерной лингвистики, возможно структурировать процессы восприятия семантики и предсказать семантическую дистанцию между словами, возникающую на биологическом уровне.

Поскольку семантика — это то, что можно выразить лишь через определенные единицы, будь то звуковые волны или символы, а значит, уже произвести некий перевод сем, предоставление семантики в зашифрованном виде машине (которая является промежуточным звеном между мысленным представлением семантики и ее распознанным представлением) усложняет ее последующий анализ. Одно из возможных решений — приближение семантической структуры уже на уровне биологических сигналов к математической, вычислительной форме. Об этом говорит и М. А. Кронгауз — по его словам, «метаязык современной лингвистической семантики во многом сложился на основе различных формальных языков математической логики» [1; 13]. Это позволит избежать промежуточных этапов в виде символов и предоставить машине прямой доступ к семантической структуре внутренней речи.

Сама речь на уровне сигналов головного мозга, не говоря уже о ее семантическом наполнении, абстрактна и представлена ассоциативными полями: «ассоциативное запоминающее устройство» — так описывают головной мозг Р. Ф. Шмидт, Ф. Ланг и М. Хекман [2; 200]. Данные исследователи также подчеркивают, что, несмотря на то, что «первично в языковой области существует доминантность левого полушария для синтаксических функциональных слов и фраз» [2; 281] (именно в левом полушарии находятся традиционно соотносимые с речевой деятельностью зоны Брока и Вернике), все же левое полушарие концентрируется «на каузальных интерференциях, на причинно-следственных отношениях и устранении логических противоречий» и задействовано «в последовательной переработке информации» [2; 282], в то время как к функциям правого полушария относятся «понимание речи; узнавание слова <…>; создание мелодии предложения и ударения (просодия); классификация речевых актов…» [2; 286]. Параллель с ассоциативными полями может быть проведена с семантическими полями, используемыми в рамках компьютерной лингвистики в качестве успешного

способа представления семантики для машины. Однако в отличие от семантических полей, выстраеваемых на основе векторов, в случае с биологическим представлением семантики более приближенным методом является учет ассоциативных рядов.

В целях оптимизации определения семантики слова разумным является сокращение семантических единиц и представление семантики объекта с помощью математических выражений. Как пишет В. Е. Булкин, «используя линейные предикатные операции, можно описывать правила образования семантических связей между соответствующими лингвистическими объектами» [3; 37], соответственно, исключается необходимость в хранении всех возможных значений единиц языка, но возникает задача конструирования семантики слова путем совершения логических операций над его составными компонентами, а также представляется важным нахождение способа выделения наиболее общих сем, чтобы отделить структурополагающие единицы от тех, что можно представить в виде комплекса базисных сем.

Исходя из концепции ограниченного числа смысловых единиц, М. А. Кронгауз выражает мысль о том, что «семантика языковых единиц описывается, как правило, не независимо, а в сопоставлении с другими языковыми единицами, как правило, близкими по значению» [1; 79]. Подобную связь языковых единиц можно встретить в толковых словарях, где для каждого слова дается определение, содержащее в себе так или иначе взаимосвязанные с означаемым словом означающие слова. Соответственно, мы можем представить определения как набор основных сем, из которых состоит семантика определяемого слова. Данными комплексами сем проще оперировать, а также их можно использовать для сравнения слов между собой — например, их близость к другим словам с семантической точки зрения и, тем самым, определение семантических полей.

Соответственно, полученные семантические поля на машинном уровне на основе базовых сем могут быть сопоставлены с семантическими полями на уровне головного мозга. Об этом, с точки зрения психолингвистики, пишет и А. Р. Лурия — «процесс восприятия слова на самом деле следует рассматривать как сложный процесс выбора нужного «ближайшего значения слова» из всего вызванного им «семантического поля»» [4; 41]. В рамках данного исследования мы фокусируемся лишь на одной из характеристик объектов, находящихся в семантическом поле, а именно — на их семантической дистанции по отношению друг к другу, которая может быть выражена через расстояние Левенштейна. Обычно оно используется при определении близости написания слов, однако в рамках нашей задачи его можно применить для сравнения закодированных комплексов сем.

Проблема исследования. Наша гипотеза заключается в том, что алгоритм представления семантики на машинном уровне, выраженный через кодирование с помощью генетического метода базовых сем слов — а именно получаемое в результате значение их семантической дистанции, высчитанной на основе расстояния Левенштейна для закодированных последовательностей базовых сем — может коррелировать с семантической дистанцией на уровне электрической активности головного мозга, зарегистрированной на ЭЭГ.

Цель нашего исследования — обучить многослойный перцептрон на базе данных ЭЭГ для его дальнейшего использования при решении задач распознавания семантики внутренней речи, а также семантической дистанции слов, закодированной с помощью предлагаемого нами алгоритма.

Для достижения цели были поставлены следующие задачи:

1. Обработать датасет, содержащий данные электрической активности мозга.
2. Закодировать семантику слов через генетическое представление их сем.
3. Вычислить семантическую дистанцию между словами и стандартизировать ее значения для последующего сопоставления результатов.
4. Обучить многослойный перцептрон и проанализировать предсказанные значения семантической дистанции на уровне сигналов мозга.

Материалы и методы. В качестве основы для обучения многослойного перцептрона был взят датасет, созданный Х. Кальво, Х. Л. Паредес Паредес и Х. Фигероа Назуно [5], содержащий данные электроэнцефалограмм, полученные при реакции участников эксперимента на предъявляемые слова-стимулы из одного или разных семантических полей. Однако для нашей работы мы переработали датасет, оставив только данные ЭЭГ при обработке слов во время третьего временного диапазона — после предъявления двух слов-стимулов.

Всего в ходе эксперимента участникам предъявлялось 138 слов-стимулов на испанском языке, каждое из которых мы соотнесли с определением для его эквивалента на английском языке для упрощения их анализа, чтобы исключить необходимость в обработке морфологических изменений слов, шире представленных в испанском языке, нежели чем в английском. В качестве источника определений использовался Кембриджский испано-английский словарь [6]. Примеры определений для слов-стимулов представлены в табл. 1.

Таблица 1

Примеры определений слов-стимулов

Слово-стимул	*Определение*
escudo	a broad piece of metal, wood carried as a protection against weapons something or someone that protects a design which is used as the symbol of the town, family see also coat of arms. a family badge or crest
plátano	the long curved fruit, yellow-skinned when ripe, of a type of very large tropical tree a type of tree with broad leaves
galleta	a crisp, sweet piece of dough baked in small flat cakes cookie a biscuit

Из определений были исключены частотные слова [7], а также редко встречающиеся (1-2 раза в рамках всех слов определений). В итоге каждому слову соответствовал набор из определенных лексических единиц.

При проведении исследования мы обращались к кодированию данных с помощью генетического метода, вычислению расстояния Левенштейна для нахождения семантической дистанции, созданию графов для визуализации семантических полей и обучению модели машинного обучения — многослойного перцептрона.

Чтобы определить семантическую дистанцию между словами, мы предлагаем кодирование наборов лексических единиц, полученных после обработки определений слов, с помощью нулей и единиц (в данном случае мы проводим параллель с генетическим методом отбора признаков, где ноль — отсутствие данной единицы в наборе, единица — его наличие). Примеры фрагментов закодированных определений представлены в табл. 2.

Таблица 2

Примеры закодированных определений слов-стимулов

Слово	*Определение*				
	Water (Вода)	*Animal (Животное)*	*Land (Суша)*	*Horse (Лошадь)*	*Sea (Море)*
Ducha (Душ)	1	0	0	0	0
Rana (Лягушка)	1	1	1	0	0
Soda (Содовая)	1	0	0	0	0
Asno (Осел)	0	1	0	1	0
Playa (Пляж)	1	0	1	0	1

Затем закодированные последовательности сравнивались между собой — выявлялась их семантическая дистанция — с помощью расстояния Левенштейна, при этом веса для каждой из трех операций — вставка, удаление и замена — были равны. Далее числовые значения расстояния использовались в качестве весов во взвешенной матрице, отражающей «координаты» каждого из слов в семантических полях, которые представлены в виде графа (рис. 1) с обозначениями на русском языке для удобства восприятия.

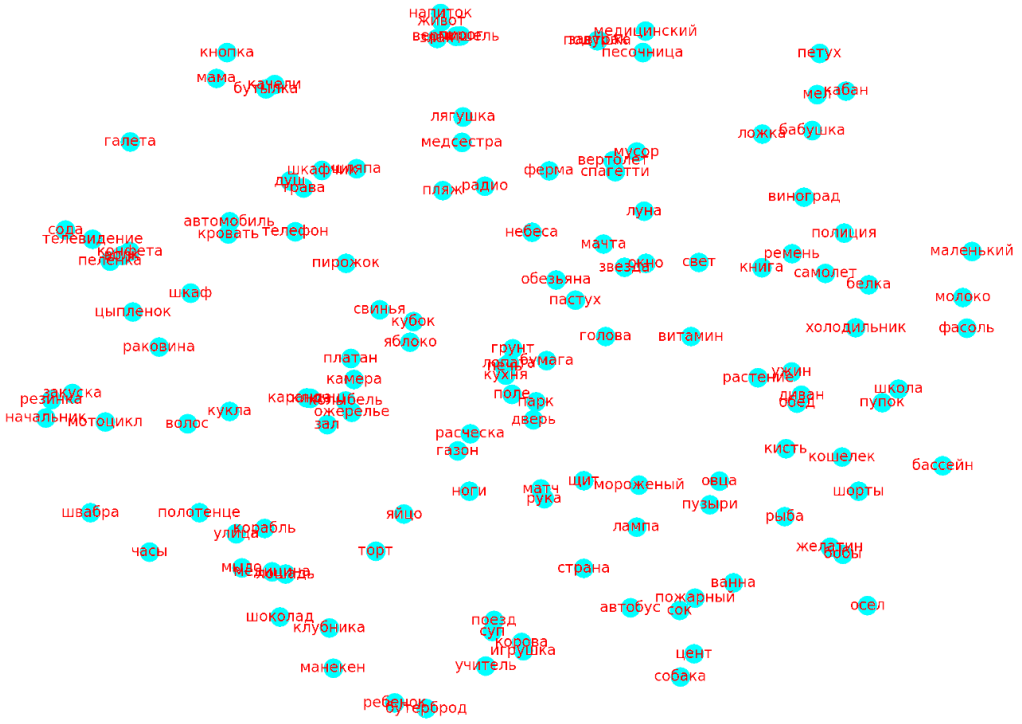

Рис. 1. Граф семантических полей

Как видно из приведенного выше графа, определенные слова находятся семантически близко, в нашем представлении, по определенному признаку, например, «лопата», «грунт», «поле», однако некоторые сочетания могут выглядеть противоречиво, как например «медицинский» и «песочница». Поскольку расположение слов в семантическом пространстве напрямую зависит от используемых определений, в дальнейшем, если будет необходимость в уве-

личении точности взаимоположения лексических единиц, возможно обращение к иным словарям. Однако в рамках нашей задачи ключевую роль играет именно сам факт возможности нахождения семантической дистанции, вычисленной по предлагаемому нами алгоритму, на основе сигналов мозга.

Чтобы выявить закономерности между компьютерным представлением семантических полей и биологическим, мы обратились к многослойному перцептрону для предсказания значений семантической дистанции у пар слов, представляемых участникам эксперимента, на основе электрической активности мозга. Для получения точности предсказания были использованы метрики MSE и RMSE.

Результаты. Обучив многослойный перцептрон на значениях электроэнцефалограмм для предсказания целевого признака — высчитанной семантической дистанции между предъявляемыми словами-стимулами в момент регистрации анализируемых в данный момент фреймов ЭЭГ, мы достигли следующих значений метрик: MSE равнялось 47,607, а RMSE — 6,9. При этом минимальное значение дистанции равнялось 0 (когда представлялись два идентичных слова), а максимальное — 38.

Заключение. Полученные значения точности позволяют прийти к выводу, что многослойный перцептрон может использоваться для предсказания семантической дистанции между словами при анализе сигналов мозга. Также, структура семантических полей на биологическом уровне коррелирует со структурой машинных семантических полей, высчитанных через предложенный нами алгоритм.

Однако в перспективе, для повышения результатов точности, возможно обращение к иным словарям для получения определений (например, составляя их самостоятельно, ориентируясь на ассоциативные ряды, или же учитывая словари синонимов), а также к изменению весов при вычислении расстояния Левенштейна.

СПИСОК ЛИТЕРАТУРЫ

1. Кронгауз М.А. Семантика / М.А. Кронгауз. — Москва: Издательский центр «Академия», 2005. — 352 с. — ISBN 5-7695-2016-7.
2. Шмидт Р.Ф. Физиология человека с основами патофизиологии: в 2 т. Т. 1. / под ред. Р.Ф. Шмидта, Ф. Ланга, М. Хекманна. — Москва: Лаборатория знаний, 2021. — 540 с.
3. Булкин В.Е. Линейные логические операторы как инструмент описания семантических правил в текстах ея / В.Е. Булкин — Текст: электронный // Вестник Херсонского национального технического университета. — 2013. — № 1 (46). — С. 36-38. — URL: https://cyberleninka.ru/article/n/lineynye-logicheskie-operatory-kak-instrument-opisaniya-semanticheskih-pravil-v-tekstah-eya (дата обращения: 11.11.2022).
4. Лурия А.Р. Язык и сознание / А.Р. Лурия; под ред. Е.Д. Хомской. — Москва: Издательство Московского университета, 1979. — 320 с.
5. Calvo H. Measuring Concept Semantic Relatedness through Common Spatial Pattern Feature Extraction on EEG Signals / H. Calvo, J.L. Paredes Paredes, J. Figueroa Nazuno // Mendeley Data. — 2018. — URL: https://data.mendeley.com/datasets/shzz5kbsgy/1 (дата обращения: 02.05.2023).
6. Cambridge Dictionary URL: https://dictionary.cambridge.org/dictionary/spanish-english/ (дата обращения: 22.05.2023).
7. Wikipedia — Most common words in English. — URL: https://en.wikipedia.org/wiki/Most_common_words_in_English (дата обращения: 22.05.2023).