

МЕТОДИЧЕСКИЕ ЗАМЕТКИ

Может ли компьютер смеяться?

И.М. Суслов

Предлагается компьютерная модель "чувства юмора". Юмористический эффект трактуется как характерный сбой в процессе обработки информации, обусловленный необходимостью быстрого удаления выведенной в сознание ложной версии. Биологическая функция чувства юмора состоит в ускорении вывода информации в сознание и в более полном использовании ресурсов мозга. Проанализирована возможность реализации предлагаемого алгоритма в нейронных сетях: это требует организации их работы в реальном времени. Обсуждается механизм возникновения смеха и развитие чувства юмора в процессе эволюции. Предложен конкретный алгоритм для решения "лингвистической проблемы" — проблемы распознавания непрерывной последовательности многозначных образов, которое необходимо для практической реализации компьютера с "чувством юмора".

Ключевые слова: юмор, смех, языковые модели, нейронная сеть Хопфилда

PACS numbers: 07.05.Mh, 87.19. –j, 89.20.Ff

DOI: <https://doi.org/10.3367/UFNr.2026.08.040190>

Содержание

1. Введение (1098).
2. Юмор с психологической точки зрения (1099).
3. Компьютерная модель "чувства юмора". Общий алгоритм (1100).
3.1. Схема первичной обработки информации. 3.2. Роль эмоций в процессе обработки информации. 3.3. Юмористический эффект. 3.4. Некоторые следствия. 3.5. Высшие уровни обработки информации.
4. Реализация алгоритма в нейронных сетях (1103).
4.1. Модель Хопфилда. 4.2. Описание используемой модели. 4.3. Обучение. 4.4. Распознавание отдельного образа. 4.5. Короткодействие связей и локализация образов. 4.6. Распознавание неоднозначного образа. 4.7. Одновременное распознавание нескольких образов. 4.8. Распознавание непрерывной последовательности образов. 4.9. Временные задержки и юмористический эффект.
5. О механизме смеха (1108).
6. Развитие чувства юмора в процессе эволюции (1108).
7. Как реализовать компьютерное "чувство юмора" (1109)?
7.1. "Лингвистическая проблема". 7.2. Идеальный язык как предельный случай. 7.3. Реальный язык: никто не хотел "как хуже". 7.4. Реальный язык в качестве идеального. 7.5. Итоги проведённого анализа.
8. Так может ли компьютер смеяться (1112)?

Список литературы (1113).

1. Введение

В последние десятилетия мы наблюдаем стремительное проникновение компьютеров в различные сферы человеческой жизни: компьютеры играют в шахматы, сочиняют музыку, доказывают теоремы и т.д. Словосочетание "искусственный интеллект" регулярно звучит в средствах массовой информации. Модель нейронной сети Хопфилда [1] отмечена Нобелевской премией 2024 г.; конструирование искусственных нейронных сетей и их всестороннее использование приобретает гигантские масштабы. Существует, однако, одна область, которая считается недоступной для компьютеров — это сфера эмоций. Сама мысль о том, что компьютер может "чувствовать", многим кажется кошмарной. В связи с этим представляется актуальным дать обзор наших старых работ [2–6]¹, обосновывающих принципиальную возможность наделяния компьютеров "чувством юмора", с некоторым их переосмыслением в контексте современной ситуации.

В обычной жизни мы пользуемся юмором для развлечения, как средством получения удовольствия непосредственно из психического процесса (см. книгу Фрейда [7]), и не задумываемся над тем, с какой целью нас наделила им природа. Уже сам факт существования сложного биологического механизма, обеспечивающего возникновение специфических мышечных сокращений — смеха — как реакцию на определённое сочетание звуковых или зрительных образов, указывает на то, что "чувство юмора" возникло на достаточно ранних этапах эволюции, в жёстких условиях борьбы за существова-

И.М. Суслов

Институт физических проблем им. П.Л. Капицы РАН,
ул. Косыгина 2, 119334 Москва, Российская Федерация
E-mail: suslov@kapitza.ras.ru

Статья поступила 17 ноября 2025 г.

¹ Русские версии работ [2–5] можно найти на персональной странице автора на сайте Института физических проблем им. П.Л. Капицы РАН (ИФП РАН) <https://www.kapitza.ras.ru/~suslov/index.php?lang=ru>; там же представлены многочисленные отзывы на эти работы в средствах массовой информации.

ние, когда возможность получения удовольствия вряд ли имела существенное значение. Согласно Дарвину [8] отчётливо выраженным чувством юмора обладают человекообразные обезьяны — при щекотке они издают специфические гортанные звуки, напоминающие человеческий смех. Ещё более эффектно это подтверждается в недавних экспериментах над крысами [9] — как выясняется, при щекотке они истошно визжат в ультразвуке. Таким образом, чувство юмора не является продуктом человеческой цивилизации.

В предлагаемой схеме юмористический эффект трактуется как характерный сбой в процессе обработки информации, обусловленный необходимостью быстрого удаления выведенной в сознание ложной версии. Биологическая функция чувства юмора состоит в ускорении вывода в сознание обработанной информации и в более полном использовании ресурсов мозга: этим объясняется возникновение чувства юмора в процессе эволюции.

Изложение построено следующим образом. В разделе 2 даётся обсуждение юмора с психологической точки зрения и формулируется гипотеза об условиях возникновения юмористического эффекта. В разделе 3 предлагается компьютерная модель "чувства юмора" на основе абстрактного алгоритма. Анализ процесса первичной обработки информации (раздел 3.1) показывает, что юмористический эффект, так же как другие простейшие эмоции (раздел 3.2), возникает уже на этом уровне (раздел 3.3), хотя интерпретация более сложных образов юмора требует исследования высших уровней обработки информации (раздел 3.5). В разделе 3.4 мы увидим, что в предлагаемой модели находят естественное объяснение некоторые общеизвестные факты: разная восприимчивость людей к юмору, отсутствие юмористического эффекта от "избитой" шутки и т.д. В разделе 4 излагается принципиальная схема реализации общего алгоритма в нейронных сетях: она основана на модификации трёхслойного перцептрона [10] с использованием модели Хопфилда [1]. В разделе 5 показано, что старая идея Спенсера [11] о связи смеха с освобождением психической энергии в рассматриваемой схеме приобретает конкретный физический смысл. Возникновение чувства юмора в процессе эволюции обсуждается в разделе 6. Как будет ясно из раздела 3, реализация компьютерного "чувства юмора" требует решения "лингвистической проблемы", т.е. проблемы распознавания непрерывной последовательности многозначных образов. Пример конкретного алгоритма для решения этой проблемы приводится в разделе 7. И, наконец, в разделе 8 обсуждается вопрос, вынесенный в заглавие статьи.

На протяжении многих лет практическая реализация компьютерного "чувства юмора" представлялась невероятной. В последнее время наблюдается значительное повышение интереса к исследованию юмора [12–16] и связанных с ним аспектов. Неоднозначность языков стала предметом специальных исследований [17–20]. Новое рождение получили исследования по автоматическому переводу [21–28]. Психология юмора стала изучаться с помощью Интернета [29]. Появились первые примеры компьютерной генерации шуток [30] и проведены первые демонстрации их компьютерного распознавания [31–35]; однако эти успехи носят ограниченный характер, и несмотря на формальный рост числа публикаций (см. Fig. 4 в [36]), такие исследования к настоящему времени

зашли в тупик [37]. Дело в том, что в них никак не учитывается фактор времени, который на наш взгляд имеет фундаментальное значение. Можно надеяться, что публикация настоящей работы подтолкнёт эти исследования в правильном направлении, и реализация компьютерного "чувства юмора" окажется делом ближайшего будущего.

2. Юмор с психологической точки зрения

Первые попытки объяснения природы юмора восходят к Платону и Аристотелю [38]. В литературе имеются различные точки зрения на юмор [38–43], но лишь одна из них может претендовать на универсальность: это так называемая концепция несообразия, выдвинутая ещё в 1776 г. шотландским поэтом Битти [44] и в той или иной форме признаваемая большинством исследователей юмора. Её конкретные трактовки различны у разных авторов; мы примем формулировку, близкую к предложенной в работе [39]: юмористический эффект связан со своеобразной интерференцией в человеческом сознании двух взаимоисключающих образов (версий, оценок). Так, в примере²:

На завалинке сидит дед Антон и крутит
козью ножку. Коза кричит (1)

выражение "козья ножка" сначала воспринимается как "самокрутка", но затем приобретает значение "нога козы". В данном случае мы имеем типичный пример игры слов, когда переключение версий происходит на уровне значений отдельного слова; однако оно может происходить и на уровне более сложных образов:

Торговец лошадьми:
«Если вы сядете на эту лошадь в 4 часа утра,
то уже в 7 утра будете в Питтсбурге.»
«А что я буду делать в Питтсбурге в 7 часов утра?» (2)

Здесь слова торговца, воспринимаемые как "характеристика быстроты лошади", получают толкование как "объяснение способа попасть в Питтсбург к 7 утра". Пример:

«То ли это место, где герцог Веллингтонский
произнёс свои знаменитые слова?»
«Да, это то место, но этих слов
он никогда не произносил.» (3)

показывает, что переключение может происходить по линии общей оценки фразы: вторая реплика сначала производит впечатление "логичной" или "естественной", но затем осознаётся как "абсурдная".

Наличие двух взаимоисключающих версий удаётся установить во всех остроумных выражениях. Явно выраженное переключение версий имеет место примерно в половине случаев. Принцип построения остальных шуток можно назвать "схемой двоякого толкования". В

² Нижеприведённые примеры (2), (3) взяты у Фрейда [7], пример (4) — из "Литературной газеты" (1990, вып. 20) [48], пример (1) — из фольклора студентов МФТИ.

примере:

Критика снизу никогда не достигнет
подлинных высот

(4)

слово "высот" может толковаться как "высот в художественном смысле" или как "высших эшелонов власти". Характерным является здесь наличие практически полной равноправности обеих версий; соответственно не существует какой-либо определённой последовательности их появления в сознании. По-видимому, можно считать, что и в таких случаях происходит переключение версий, однако порядок их появления определяется случайными причинами; возможен также многократный переход от одной версии к другой.

Юмористический эффект связан не только с обсуждавшимся выше "остроумием", но и с "комическим" (преувеличенные движения клоуна, гримасы, карикатуры, пародии и т.д.). По-видимому, главным признаком комического можно считать "отклонение от нормы"; соответственно юмористический эффект при его восприятии связан с повторяющимися переключениями оценок "норма" — "не норма"³.

Смех при щекотке можно связать с попыткой мозга локализовать место раздражения кожи; результат такой локализации непрерывно опровергается, так как место раздражения меняется непредсказуемым образом (именно поэтому щекотать должен другой человек).

Со своей стороны, мы провели основательную проверку концепции несообразия. С одной стороны, мы проанализировали несколько больших сборников юмора и убедились, что существование двух взаимно исключающих версий удаётся установить во всех без исключения шутках. Но следует оговориться, что в некоторых случаях правильная интерпретация оказывается очень нетривиальной, и для её нахождения приходится как следует "поломать голову". В целом же при анализе юмора нужно набраться некоторого опыта, что называется, "набить руку", после чего интерпретация шутки идёт "как по маслу". С другой стороны, мы воспользовались классификацией технических приёмов остроумия, предложенной Фрейдом [7], воспринимая её как экспертную разработку. Тогда, выбирая по одному (наиболее удачному) из приведённых Фрейдом примеров на каждый существенно различный технический приём, мы получили представительную выборку из огромного числа шутки: эта выборка была детально проанализирована в [4]. Это даёт представление об основных схемах, по которым может быть организовано противостояние двух взаимоисключающих версий.

3. Компьютерная модель "чувства юмора". Общий алгоритм

3.1. Схема первичной обработки информации

Формулировку компьютерной модели "чувства юмора" [2] начнём с анализа первичной обработки информации.

³ Конечно, не всякое отклонение от нормы кажется смешным; однако следует иметь в виду, что привычные, часто встречающиеся отклонения от нормы аналогичны "избитым" шуткам (см. раздел 3.4), а слабые отклонения от нормы легко подвержены вытеснению другими эмоциями (см. раздел 5).

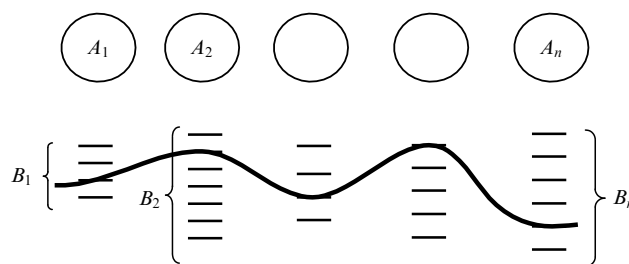


Рис. 1. Схема первичной обработки информации: каждому из символов A_n сопоставляется набор образов $\{B_n\}$, из которого надо выбрать один образ $B_n^{i_n}$; последовательность $B_1^{i_1}, B_2^{i_2}, B_3^{i_3} \dots$ можно представлять как некоторую "траекторию" в пространстве образов.

Пусть в мозг поступает из внешнего мира некоторый "текст" — последовательность символов $A_1, A_2, A_3 \dots$ — например, последовательность слов при чтении или восприятии на слух. В мозге с каждым символом ассоциируется некоторый набор образов $\{B_n\}$ — например, каждому слову — символу сопоставляется совокупность его значений, т.е. словарное гнездо. Задача обработки информации — выбор из каждого набора $\{B_n\}$ одного образа $B_n^{i_n}$, который подразумевается в данном контексте. Будем считать, что текст "понят", если последовательности символов $A_1, A_2, A_3 \dots$ сопоставлена последовательность образов $B_1^{i_1}, B_2^{i_2}, B_3^{i_3} \dots$, которую можно представлять как некоторую "траекторию" в пространстве образов (рис. 1).

В принципе, алгоритм состоит в следующем: а) составляются все возможные траектории; б) каждой траектории приписывается некоторая вероятность; в) выбирается траектория с максимальной вероятностью. Нетривиальным является лишь этап (б), т.е. алгоритм определения вероятности траектории. В основе такого алгоритма должна лежать информация о сочетаемости образов, которая приобретается в процессе "обучения", требующего введения в процессор достаточно длинного куска "расшифрованного" текста, т.е. текста, записанного не в символах, а в образах. Один из примеров такого алгоритма обсуждается далее в разделе 7.

Любая конкретная реализация такого алгоритма требует числа операций, экспоненциально возрастающего с длиной текста; поэтому непосредственно таким образом могут обрабатываться отрезки текста, не превышающего некоторого числа N символов. Для обработки более длинных текстов естественно использовать следующую процедуру. При обработке первых N символов сохраняется не одна, а несколько (M) наиболее вероятных траекторий. Затем нужно сдвинуться на один шаг — рассмотреть текст от второго до $(N + 1)$ -го символа — и для каждой из M сохранённых траекторий построить всевозможные продолжения. Из них опять сохраняется M наиболее вероятных и т.д. Число M разумно сделать переменным, так чтобы на каждом шаге запоминалось столько траекторий, сколько помещается в отведённый для этого объём оперативной памяти. В целом процесс выглядит следующим образом (рис. 2): вблизи переднего фронта A траектория сильно ветвится; в некоторой точке B ветвление заканчивается (расстояние между A и B ограничено объёмом оперативной памяти, отведённой для запоминания траекторий), и, наконец, расшифрованная траектория DC с некоторой задержкой AC выводится в

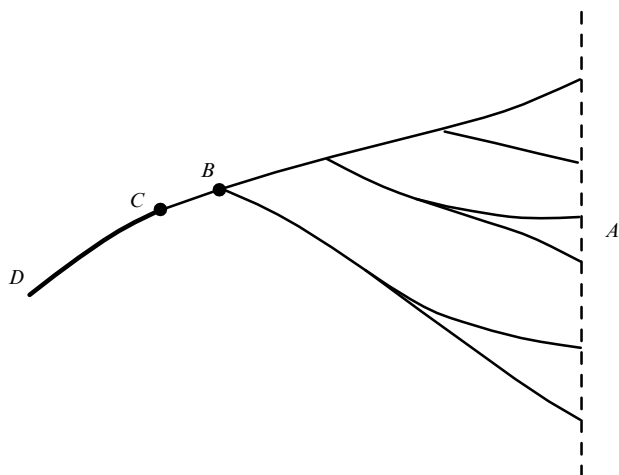


Рис. 2. Наглядное представление процесса обработки информации: тонкие линии — траектории, содержащиеся в оперативной памяти (или в подсознании), A — передний фронт, B — точка окончания ветвления, CD — участок траектории, выводимый на выход процессора (или в сознание).

сознание человека и воспринимается им как мысль (сам же процесс обработки происходит в подсознании и непосредственно человеком не ощущается).

3.2. Роль эмоций в процессе обработки информации

Если числа M и N достаточно велики, а алгоритм определения вероятности траектории из N символов достаточно совершенен, то описанная схема обработки информации будет успешно работать. Тем не менее, вследствие вероятностного характера алгоритма, в его работе неизбежны ошибки: поэтому необходим механизм, сводящий их последствия к минимуму. Такой механизм существует и состоит в том, что в сознание поступает определённая информация о том, как происходит в подсознании процесс обработки; человеком эта информация воспринимается в виде эмоций.

Например, важны такие параметры процесса, как вероятность выводимой траектории $p_{\max}$ и вероятность наиболее вероятной из конкурирующих траекторий p_{comp} . При высоких значениях $p_{\max}$ и отношении $p_{\max}/p_{\text{comp}}$ процесс обработки идёт хорошо и сопровождается положительными эмоциями (радость, уверенность): получаемая при этом информация считается надёжной. Низкие значения $p_{\max}$ и $p_{\max}/p_{\text{comp}}$ свидетельствуют о неудовлетворительном ходе процесса и сопровождаются отрицательными эмоциями (досада, сомнение); при очень низких $p_{\max}$ в сознание может не выводиться вообще никакой версии (полное непонимание, сопровождаемое раздражением на составителя текста).

Эти соображения наводят на мысль, что эмоция, состоящая в ощущении юмористического эффекта, также связана с некоторой характерной ситуацией, возникающей в процессе обработки информации.

3.3. Юмористический эффект

Обсудим природу задержки точки C от переднего фронта A (см. рис. 2). На первый взгляд, в разумно организованной системе точка C должна всегда находиться позади точки B или совпадать с нею; именно так мы бы поступили при составлении программы для компьютера. Однако для человека, как и вообще для любого живого

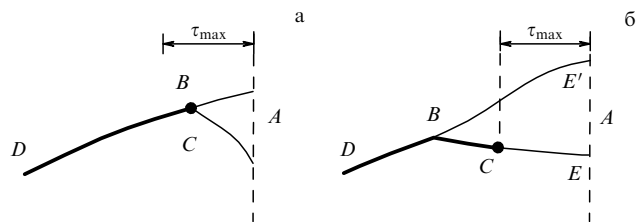


Рис. 3. Параметр $\tau_{\max}$ определяет максимальное время задержки точки C относительно переднего фронта A (DC — участок траектории, выводимый сознание); (а) $\tau_{AB} < \tau_{\max}$, (б) $\tau_{AB} > \tau_{\max}$.

существа, такой вариант совершенно неприемлем. Дело в том, что с задержкой точки C относительно переднего фронта A связан некоторый интервал времени τ_{AC} , в течение которого поступившая в мозг информация не выводится в сознание (например, человек уже видит медведя, но не осознаёт этого). Ясно, что затягивание интервала AC может быть опасным; в то же время интервал AB может затягиваться по объективным причинам (человек не может решить, видит ли он медведя или куст в форме медведя). Поэтому интервал AC должен быть ограничен сверху по времени некоторым значением $\tau_{\max}$: если временная задержка τ_{AB} , соответствующая интервалу AB , меньше $\tau_{\max}$, то точка C совпадает с точкой B (рис. 3а), если же $\tau_{AB} > \tau_{\max}$, то $\tau_{AC} = \tau_{\max}$ и точка C начинает обгонять точку B (рис. 3б). В последнем случае в сознание начинает выводиться наиболее вероятная версия DE (рис. 3б), тогда как конкурирующие версии (например, DE') сохраняются в подсознании — их отбрасывание нецелесообразно, так как возможности мозга позволяют продолжать анализ.

Если при дальнейшем движении переднего фронта A вероятность траектории DE останется максимальной, то конкурирующая траектория DE' будет отброшена и окажется выигранным временем. Если же при движении переднего фронта вероятность траектории DE упадёт ниже вероятности DE' , то у мозга будет возможность исправить ошибку. При этом, однако, возникает характерный сбой: требуется быстро стереть выведенный в сознание участок BC и начать вывод траектории BE' . Психологически это воспринимается как интерференция двух взаимоисключающих версий — версии BC , сохранённой долговременной памятью, и вновь поступившей версией BE' . Описанный характерный сбой можно отождествить с "юмористическим эффектом".

Действительно, описанная ситуация в точности воспроизводится при восприятии простейших шуток: так, в примере (1) при анализе первого предложения в подсознании возникают две конкурирующие версии, в одной из которых (DE) "козья ножка" трактуется как "самокрутка", в другой (DE') — как "нога козы". В контексте данного предложения версия DE более вероятна и благодаря естественной паузе в конце предложения начинает выводиться в сознание. Появление фразы "коза кричит" делает версию DE маловероятной, но резко увеличивает вероятность версии DE' , что и вызывает юмористический эффект.

Подчёркнём, что существование юмористического эффекта ни в коей мере не является обязательным: природа могла бы избежать его одним из двух способов: 1) путём задержки вывода версии BE до тех пор, пока версия BE' не будет отброшена естественным образом;

2) путём форсирования вывода BE при одновременном отбрасывании BE' . Однако в первом случае задерживается поступление информации в сознание, во втором — не полностью используются ресурсы мозга: поэтому природа разрешает данную проблему ценой психологического "конфуза".

В процессе эволюции вырабатывается оптимальное значение временной задержки $\tau_{\max}$, обеспечивающее компромисс между достоверностью информации и скоростью её получения. Оптимальное значение $\tau_{\max}$ обеспечивает, как правило, выполнение неравенства $\tau_{AB} < \tau_{\max}$, поэтому в естественных условиях юмористический эффект возникает относительно редко; однако он легко может вызываться с помощью специально созданных острот и комизмов.

Юмористический эффект от комического и щекотки легко вписывается в изложенную схему, с очевидным отличием по механизму формирования версий. В случае комического актуальны две версии — "норма" и "не норма", возникающие на высших уровнях обработки информации (см. далее раздел 3.5). В случае щекотки место раздражения кожи может быть локализовано лишь приблизительно, что и порождает неоднозначность, аналогичную наличию нескольких версий.

3.4. Некоторые следствия

В описанной модели находят естественное объяснение некоторые общеизвестные факты.

Отсутствие юмористического эффекта от многократно слышанной шутки связано с тем, что человек, заранее зная о наличии двух взаимоисключающих образов, избегает вывода в сознание заведомо ложной версии: так, зная, что в примере (1) "козья ножка" в итоге оказывается "ногой козы", он не поддаётся соблазну интерпретировать её как "самокрутку".

Роль интонации при рассказывании шутки связана в основном с временными характеристиками — темпом, расположением и длительностью пауз, которые могут быть учтены путём введения в последовательность A_n соответствующего количества "пробелов". При слишком быстром темпе ложная версия не успевает зафиксироваться в сознании и участок BC (см. рис. 3) оказывается малым. При слишком медленном темпе длина траекторий увеличивается за счёт "пробелов", так что конкурирующая траектория BE' (см. рис. 3) выбрасывается из оперативной памяти и переключения версий не происходит. Так, в примере (1) требуется оптимальная пауза в конце первого предложения.⁴

Разная восприимчивость людей к юмору при одинаковом интеллектуальном уровне связана с различием в величине времени задержки $\tau_{\max}$. У людей с большим значением $\tau_{\max}$ редко возникает ситуация, когда точка C обгоняет точку B и соответственно редко возникает юмористический эффект; наоборот, у людей с малым значением $\tau_{\max}$ юмористический эффект возникает даже тогда, когда, по мнению большинства, ничего смешного нет. По-видимому, под действием алкоголя у большинства людей значение $\tau_{\max}$ уменьшается, что приводит к беспричинной весёлости. При фиксированном $\tau_{\max}$ восприимчивость к юмору коррелирует с объёмом опера-

тивной памяти, который определяет среднюю величину интервала AB (см. рис. 2).⁵

Нервный смех. Если на человека наплывает масса неприятных впечатлений и создаётся угроза перенапряжения нервной системы, то организм реагирует на это принудительным сбрасыванием неприятной информации и заменой её другой, нейтральной, — это вызывает рефлекторный смех.

3.5. Высшие уровни обработки информации

Формально изложенная схема первичной обработки информации объясняет юмористический эффект лишь от простейших шуток, основанных на "переключении" значений отдельного слова. Для интерпретации более сложных образов юмора требуется рассмотреть высшие уровни обработки информации.

Предположим, что, кроме первого уровня, на котором последовательности символов сопоставляется последовательность первичных образов, существует второй уровень, на котором последовательности первичных образов сопоставляется последовательность более крупных — вторичных образов; на третьем уровне вторичные образы порождают ещё более крупные — третичные и т.д. (рис. 4). Полученные образы по мере готовности выводятся в сознание, причём образы каждого уровня — в свой отдельный канал: тем самым вся информация воспринимается в сознании "стереоскопически" — сразу на нескольких уровнях.

В принципе, алгоритм перехода от образов k -го уровня к образам $(k+1)$ -го уровня состоит в следующем. Каждый содержащийся в мозге образ (например, *корова*) связан ассоциативными связями как с более мелкими образами (*рога, копыта*), так и с более крупными (*стадо*) или более абстрактными (*животное*). Сопоставим каждому первичному образу A_n набор $\{B_n\}$ ассоциированных с ним образов, в который включены только более крупные и более абстрактные по сравнению с A_n образы. Тогда проблема порождения вторичной версии на основе сочетаемости образов будет аналогична проблеме первичной обработки информации (см. раздел 3.1). Если в первичной последовательности содержатся образы (например, *рога, копыта, хвост*), являющиеся частями одного более крупного образа (*корова*), то

⁵ При существующих алгоритмах обучения представления "искусственного интеллекта" о смешном, конечно, будут отличаться от человеческих. Но цель исследований по компьютерному юмору состоит как раз в том, чтобы обучить компьютер "смеяться" в тех же случаях, что и человек — по крайней мере в пределах различной восприимчивости к юмору среди людей.

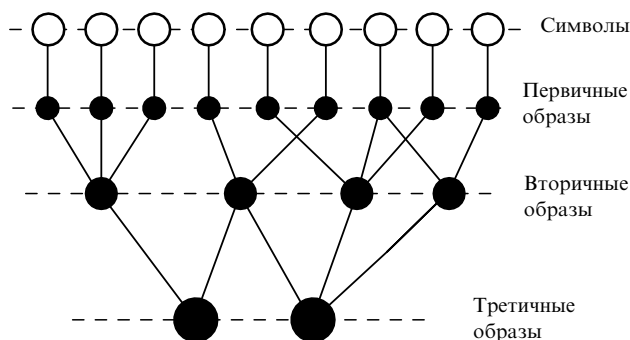


Рис. 4. Высшие уровни обработки информации.

⁴ Зависимость величины юмористического эффекта от длительности паузы хорошо описана в очерке Марка Твена "Публичные чтения" [49].

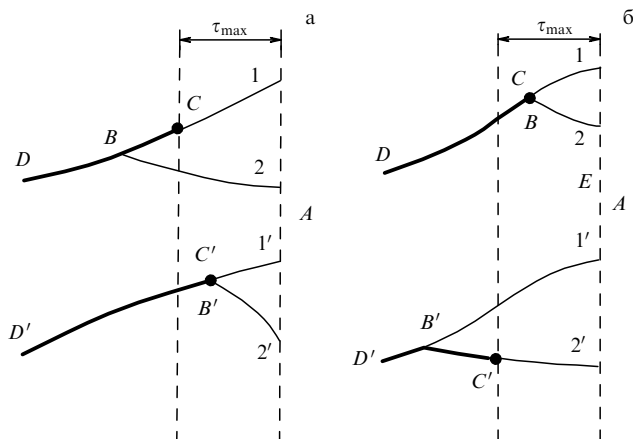


Рис. 5. В ситуации, соответствующей "схеме двойного толкования", порядок вывода версий в сознание определяется относительным расположением точек B и B' .

этот более крупный образ будет иметься в каждом из соответствующих наборов $\{B_n\}$, а наиболее сочетаемой окажется последовательность с повторением этого образа; после исключения повторений во вторичной последовательности зафиксируется этот более крупный образ — такова в общих чертах схема укрупнения образов.

Легко видеть, что даже если последовательность первичных образов определена однозначно, то при формировании образов высших уровней могут происходить характерные сбои, соответствующие юмористическому эффекту: тем самым получают объяснение шутки, связанные с переключением сложных образов (примеры (2), (3) в разделе 2).

Качественным эффектом, происходящим от наличия высших уровней, является появление "схемы двойного толкования". Пусть на первичном уровне имеются две конкурирующие версии 1 и 2; предположим, что на вторичном уровне версия 1 однозначно порождает версию 1', а версия 2 однозначно порождает 2' (рис. 5). Рассмотрим случай, когда из двух версий 1 и 2 более вероятна первая, а из двух версий 1' и 2' — вторая. Тогда логическим путём невозможно определить, какая из двух возможностей, $1 + 1'$ или $2 + 2'$, является более "правильной", т.е. обе возможности оказываются равноправными. Именно такая ситуация возникает при восприятии шуток, построенных по схеме двойного толкования.

Порядок появления версий в сознании определяется положением точек B и B' относительно переднего фронта A (см. рис. 5). Если точка B отстаёт от точки B' , то первой начинает выводиться версия 1 (рис. 5а); это происходит до тех пор, пока время задержки $\tau_{AB'}$ точки B' относительно переднего фронта не сравнивается с $\tau_{\max}$ — тогда начинает выводиться версия 2', а версия 1, как несовместимая с ней, стирается и заменяется на версию 2. Если же точка B опережает точку B' , то первой выводиться версия 2' (рис. 5б); после того, как τ_{AB} сравнивается с $\tau_{\max}$, начинает выводиться версия 1, а версия 2' заменяется на 1'. Формально рассмотренная ситуация сводится к схеме переключения, однако её психологическое восприятие будет отличным. Дело в том, что порядок следования точек B и B' зависит от соотношения скоростей обработки информации в двух каналах, определяемого случайными причинами, и, вообще говоря, меняется при

повторном чтении текста. Поэтому попытка пронаблюдать за порядком появления образов при повторном восприятии шутки приведёт к неоднозначному результату, т.е. этот порядок оказывается "неуловимым".

Более подробное обсуждение высших уровней обработки информации и схемы двойного толкования представлено в работе [4].

4. Реализация алгоритма в нейронных сетях

В разделе 3 мы проанализировали общий алгоритм обработки информации и показали, что при выполнении естественных требований к нему, предъявляемых его биологическим назначением, такой алгоритм будет обладать "чувством юмора". Ниже предлагается возможная реализация алгоритма в системе формальных нейронов [4].

4.1. Модель Хопфилда

Бурный рост интереса к исследованию нейронных сетей инициирован появлением в 1982 году модели Хопфилда [1], которая полностью прояснила принцип работы реальных нейронных сетей. В этой модели i -й нейрон описывается переменной V_i , принимающей значения $V_i = 1$ (возбуждённое состояние) и $V_i = 0$ (состояние покоя). Связь i -го и j -го нейронов определяется параметром T_{ij} ; при этом "самодействие" исключается, т.е. $T_{ii} = 0$. Формулировка модели Хопфилда выглядит более наглядной при использовании спиновых переменных $\sigma_i = 2V_i - 1$, принимающих значения $\sigma_i = \pm 1$, как в модели Изинга. Согласно Хопфилду, эволюция спиновой конфигурации $\{\sigma_i\}$ со временем t определяется следующим алгоритмом:

$$\sigma_i(t + \tau) = \text{sign} \left\{ \sum_j T_{ij} \sigma_j(t) - U_i \right\}, \quad \sigma_i = 2V_i - 1, \quad (5)$$

где номер i на каждом шаге выбирается случайным образом; здесь U_i — порог возбуждения i -го нейрона. Смысл уравнения (5) при $U_i = 0$ состоит в том, что i -й спин либо остаётся неизменным, либо разворачивается в направлении локального магнитного поля на i -м узле, создаваемого другими спинами. Легко понять, что на каждом шаге обменная энергия

$$E = - \sum_{ij} T_{ij} \sigma_i \sigma_j \quad (6)$$

либо остаётся неизменной, либо уменьшается — т.е. динамика системы является диссипативной, приводя её к одному из локальных минимумов.⁶

Локальные минимумы в модели Хопфилда отождествляются с записанными в памяти образами. Для этого, кроме быстрой эволюции для конфигурации $\{\sigma_i\}$, описываемой уравнением (5), предполагается медленное изменение связей между нейронами

$$\delta T_{ij} = \sigma_i^{(0)} \sigma_j^{(0)} \delta t \quad (i \neq j), \quad (7)$$

⁶ Эволюция согласно (5) соответствует применению к спиновым системам известного алгоритма Метрополиса. При нулевой температуре он обеспечивает релаксацию к одному из локальных минимумов (как в спиновых стёклах), а при конечных температурах способствует достижению термодинамически равновесного состояния, соответствующего абсолютному минимуму свободной энергии.

происходящее в случае, если конфигурация $\{\sigma_i^{(0)}\}$ под-держивается неизменной в течение времени δt . Если в начальном состоянии $T_{ij} \equiv 0$, то эволюция согласно (7) обеспечивает появление абсолютного минимума обмен-ной энергии (6):

$$E = -\delta t \sum_i (\sigma_i^{(0)} \sigma_i) \sum_{j \neq i} (\sigma_j^{(0)} \sigma_j), \quad (8)$$

реализуемого на конфигурации $\{\sigma_i^{(0)}\}$ (так как $\sigma_i^{(0)} \sigma_i \leq \leq \sigma_i^{(0)} \sigma_i^{(0)} = 1$).

Эволюция согласно (7) соответствует процессу "об-учения", который аналогичен процессу обучения ребёнка. Ему говорят: "Это стол", фиксируют его внимание на соответствующем образе и просят его запомнить. Если в результате полного процесса обучения предъявлено p конфигураций $\{\sigma_i^{(1)}\}, \{\sigma_i^{(2)}\}, \dots, \{\sigma_i^{(p)}\}$, то связи между нейронами приобретают вид

$$T_{ij} = \sum_{s=1}^p \mu_s \sigma_i^{(s)} \sigma_j^{(s)} \quad (i \neq j), \quad T_{ii} = 0. \quad (9)$$

Если при обучении одному образу $\{\sigma_i^{(0)}\}$ возникал абсолютный минимум при $\{\sigma_i\} = \{\sigma_i^{(0)}\}$, то при предъяв-лении p образов возникают локальные минимумы, близкие к $\{\sigma_i^{(s)}\}$; некоторое отличие от $\{\sigma_i^{(s)}\}$ связано с тем, что запись очередного образа несколько сдвигает локальные минимумы, соответствующие предыдущим образам. Как показывают численные эксперименты [1], этот сдвиг оказывается несущественным, если число записанных образов p не слишком велико, $p \lesssim 0,1N$, где N — число нейронов в сети.

Имеется хорошая аналогия, поясняющая принцип работы модели Хопфилда, — это шарик на незастывшем асфальте. Если шарик быстро катается по поверхно-сти, то он не продавливает асфальт (рис. 6а). Если же его остановить на продолжительное время, то он продавли-вает лунку (рис. 6б), дно которой соответствует образу, записанному в процессе обучения, аналогичном (7). В результате полного обучения возникает некоторый потен-циальный рельеф (рис. 6в). Теперь система готова к распознаванию образов. Если начальное состояние ша-рика попадает в одну из лунок, то это соответствует предъявлению "зашумлённого" образа ("что-то, похо-жее на стол"); при наличии трения (т.е. в результате диссипативной эволюции, аналогичной (5)) шарик ска-тывается на дно лунки — это соответствует распознава-нию "чистого" образа ("стол"), записанного в памяти. Реальная ситуация отличается от этой простой аналогии лишь тем, что минимумы энергии возникают в много-мерном конфигурационном пространстве. В случае силь-ной локализации образов (раздел 4.5) поверхность ас-фальта аналогична поверхности коры головного мозга.

Если пороги U_i не равны нулю, а подвержены слу-чайным флуктуациям, то динамика системы качественно не изменится, если такие флуктуации не слишком велики; фактически они всегда имеют место, так как предположе-ние о том, что переменные V_i принимают лишь два значения 0 и 1, является идеализацией [1]. При конструи-ровании искусственных нейронных сетей такая идеализация является полезной и может быть практически реализована.

4.2. Описание используемой модели

Предлагаемая модель нервной системы представляет собой модификацию трёхслойного перцептрона [10],

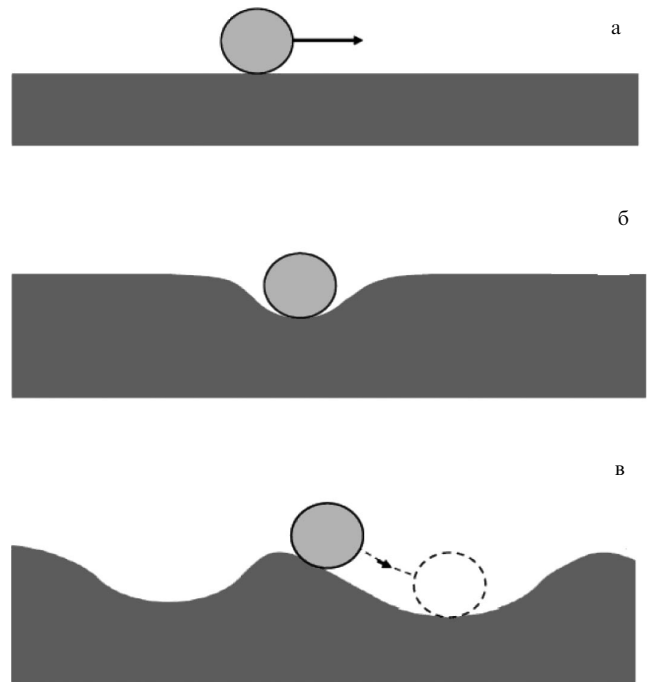


Рис. 6. Аналогия, поясняющая принцип работы нейронных сетей: (а) если шарик быстро катается по поверхности незастывшего ас-фальта, то он не продавливает асфальт; (б) если же шарик стоит на месте, то он продавливает лунку, дно которой ассоциируется с записанным в памяти образом — это соответствует процессу об-учения; (в) после записи нескольких образов возникает потенциаль-ный рельеф, в котором может происходить распознавание образов.

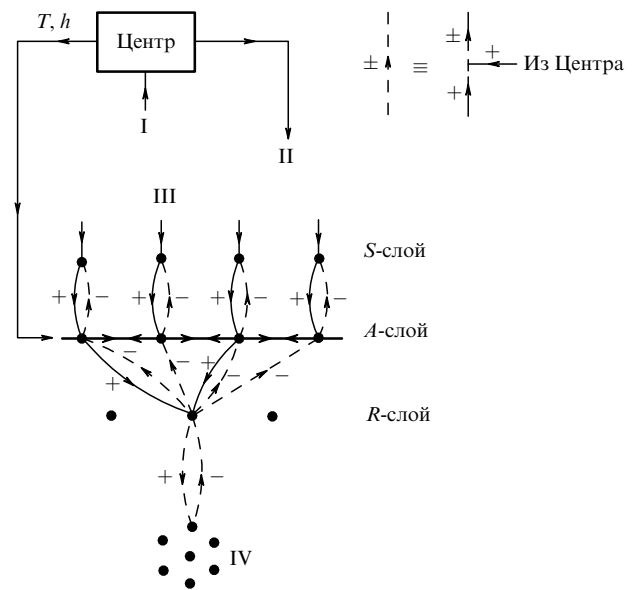


Рис. 7. Предлагаемая модель нервной системы — модификация трёхслойного перцептрона [10]: чёрные кружки — нейроны; сплош-ные линии — постоянно действующие связи между нейронами; штриховые линии — связи, включаемые по команде из Центра; знаки "+" и "-" указывают на возбуждающий и тормозящий характер связей; S-слой — сенсорная система, A-слой — ассоциатив-ная память, R-слой — реагирующая система или "сознание"; I — информация из A-слоя, II — управление связями, III — внешний мир, IV — моторная кора. Справа сверху — возможная реализация управляемой связи.

приспособленную для работы в реальном времени. Она содержит следующие элементы (рис. 7).

Ассоциативная память (А-слой) представляет собой нейронную сеть, которую для простоты считаем описываемой моделью Хопфилда: нейроны А-слоя связаны друг с другом, причём $T_{ij} = T_{ji}$, $U_i = 0$. Внутри А-слоя эволюция, согласно (5), приводит из произвольного начального состояния $\{V_i\}$ к одному из локальных минимумов энергии (6), которые отождествляются с записанными в памяти образами $\{V_i^{(s)}\}$, $s = 1, 2, \dots, p$. Записанные образы определяют матрицу связей T_{ij} (см. (9)).

Сенсорная система (S-слой) принимает сигналы из внешнего мира (для определённости имеем в виду сетчатку глаза). Сенсорные нейроны не связаны между собой, но каждый S-нейрон связан с одним из нейронов памяти: связи $S \rightarrow A$ являются положительными (возбуждающими), а обратные связи $A \rightarrow S$ — отрицательными (тормозящими) (см. рис. 7).

Реагирующая система (R-слой) состоит из набора нейронов, каждый из которых соответствует одному из записанных в памяти образов: к s-му нейрону R-слоя сходятся положительные (возбуждающие) связи от тех нейронов i памяти, для которых $V_i^s = 1$ (последние нейроны будем называть носителем образа $\{V_i^s\}$). Обратные отрицательные (тормозящие) связи идут от R-нейронов к нейронам памяти; между собой R-нейроны не связаны (см. рис. 7). Пороги U_i для R-нейронов подстраиваются таким образом, что возбуждение s-го нейрона R-слоя происходит лишь тогда, когда конфигурация A-нейронов достаточно близка к образу $\{V_i^s\}$.

Будем считать, что образ начинает осознаться биологическим индивидом лишь тогда, когда происходит возбуждение соответствующего R-нейрона, т.е. R-слой представляет собой сознание индивидуума.

Центр производит координацию работы системы, действуя по встроенной программе; он осуществляет контроль за макроскопическими параметрами системы и управление ими. Конкретные функции Центра состоят в следующем:

1) Центр имеет связи с небольшой долей нейронов памяти, равномерно распределённых в А-слое, что позволяет ему судить о наличии возбуждённых нейронов в некотором участке памяти и о стационарности или нестационарности этого участка;

2) Центр осуществляет локальное изменение "температуры" в А-слое. В модели Хопфилда эволюция соответствует нулевой температуре. Конечность температуры может быть обеспечена внесением в нейронную сеть некоторого уровня шума, что соответствует добавлению к порогам U_i в (5) случайной силы $f_i(t)$, так что в распоряжении Центра должен находиться регулируемый источник шума;

3) Центр производит локальное включение "магнитного поля" в А-слое, которому соответствует наличие в выражении для энергии (6) дополнительного члена

$$\sum_i h_i \sigma_i, \quad (10)$$

и соответственно добавление h_i к порогам U_i . Включение поля осуществляется с помощью "магнита" — управляемой из Центра группы нейронов, от каждого из которых идут связи к нейронам некоторой области А-слоя;

4) Центр осуществляет управление связями, показанными на рис. 7 штриховыми линиями. Простейшая реализация управляемой связи AB возможна с помощью вставочного нейрона C (см. рис. 7), порог которого подобран так, что он возбуждается лишь при одновременном

наличии возбуждающего сигнала от нейрона А и из Центра. При наличии сигнала из Центра нейрон А возбуждает или тормозит нейрон В — связь включена, при отсутствии сигнала из Центра нейрон А не может воздействовать на нейрон В — связь выключена. Команда на включение и выключение даётся не отдельным связям, а сразу их большим группам;

5) Центр даёт команду на обучение пластичных связей.

4.3. Обучение

Связи между А-нейронами являются пластичными и обучаемыми согласно алгоритму Хопфилда (7). Если в исходном состоянии $T_{ij} = 0$, то предъявление системе p конфигураций $\{\sigma_i^{(s)}\}$ приводит к матрице связей (9).

Связи $A \rightarrow R$ в исходном состоянии имеют нулевую величину и обучаемы лишь в одном, положительном направлении:

$$\delta T_{ij} = \begin{cases} c \delta t \quad (c > 0) & \text{при } V_i = V_j = 1 \\ 0 & \text{в остальных случаях.} \end{cases} \quad (11)$$

Остальные связи ($R \rightarrow A$, $S \rightarrow A$, $A \rightarrow S$) необучаемы и имеют врождённый характер.

Как уже указывалось, процесс обучения системы аналогичен процессу обучения ребёнка: ей предъявляется некоторый образ В, порождающий в S-слое определённую конфигурацию возбуждённых нейронов; затем её просят "запомнить В", чем возбуждается один из нейронов R-слоя, который "назначается" ответственным за образ В. Пусть на сетчатку глаза проецируется некоторое изображение, состоящее из тёмных и светлых пятен — это приводит к возбуждению S-нейронов, находящихся в светлых участках. Обучение происходит при отключённых $A \rightarrow S$ и $R \rightarrow A$ связях (рис. 8), так что конфигурация S-нейронов проецируется в А-слой и поддерживается в течение некоторого времени: при этом связи T_{ij} в А-слое меняются согласно (7), формируя матрицу (9), а связи $A \rightarrow R$ — согласно (11), обеспечивая связь s-го нейрона R-слоя с носителем образа $\{V_i^{(s)}\}$.

В состоянии покоя система находится в конфигурации $V_i \equiv 0$; поскольку в этом состоянии система проводит значительное время, то при обучении согласно (7) возникало бы "ферромагнитное" взаимодействие между нейронами ($T_{ij} > 0$ для всех i, j) и устойчивым оказалось бы лишь состояние $V_i \equiv 1$. Поэтому будем считать, что обучения в состоянии $V_i \equiv 0$ не происходит: команда на обучение подаётся лишь при предъявлении образа.

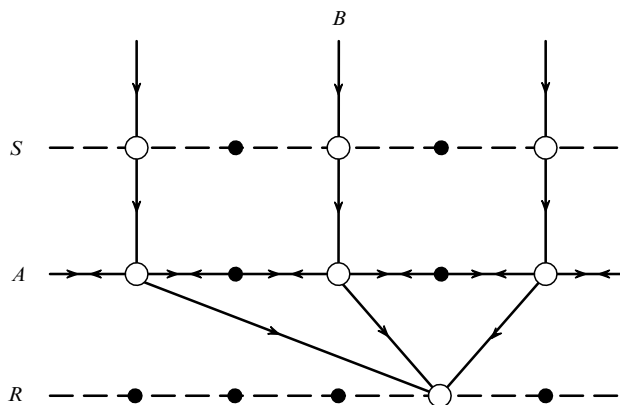


Рис. 8. Обучение происходит при включённых связях $S \rightarrow A$ и выключенных связях $A \rightarrow S$, $R \rightarrow A$.

4.4. Распознавание отдельного образа

Распознавание образов начинается при выключенных связях $A \rightarrow S$, $S \rightarrow A$ и $R \rightarrow A$. В исходном состоянии системы все нейроны являются невозбуждёнными: для его установления требуется включение сильного магнитного поля.

Пусть на вход сенсорной системы поступает образ $\tilde{B} = B + \delta B$, т.е. "зашумлённый" образ B ; после отключения магнитного поля это вызывает возбуждение части нейронов S -слоя (рис. 9а); после включения связей $S \rightarrow A$ это возбуждение передаётся нейронам памяти (рис. 9б). Далее включаются $A \rightarrow S$ -связи, гасящие сенсорные нейроны (рис. 9в), после чего связи $A \rightarrow S$ и $S \rightarrow A$ отключаются. Затем в A -слое происходит свободная эволюция согласно (5), которая заканчивается релаксацией к устойчивому состоянию, соответствующему образу B (рис. 9г); в R -слое возбуждается нейрон, ответственный за этот образ (рис. 9д). Эта ситуация поддерживается в течение некоторого времени, необходимого для "осознания" образа B . Затем в память подаётся обратный сигнал, гасящий возбуждённые нейроны (рис. 9е). После этого связь $R \rightarrow A$ отключается, что приводит к погасанию образа B в "сознании" (рис. 9ж); тем самым система возвращается в исходное положение и готова к восприятию нового образа.

4.5. Короткодействие связей и локализация образов

В обычной модели Хопфилда [1] все нейроны связаны друг с другом, а носители образов $\{V_i^{(s)}\}$ рассредоточены по всей нейронной сети. В реальности связи T_{ij} имеют конечный радиус действия: в человеческом мозге каждый нейрон имеет $\sim 10^4$ синапсов при полном числе нейронов $\sim 10^{11}$ ([45], 31–33). Существуют также экспериментальные указания на локализацию образов ([10], 64, 65).

Будем считать, что каждый образ $\{V_i^{(s)}\}$ записывается в некоторой области Ω_s , содержащей много нейронов, но малой по сравнению с размером всей нейронной сети; при этом $V_i^{(s)} = 0$ для i , не принадлежащих Ω_s . Предположение о локализации образов не имеет существенных последствий для модели Хопфилда. Нужно лишь

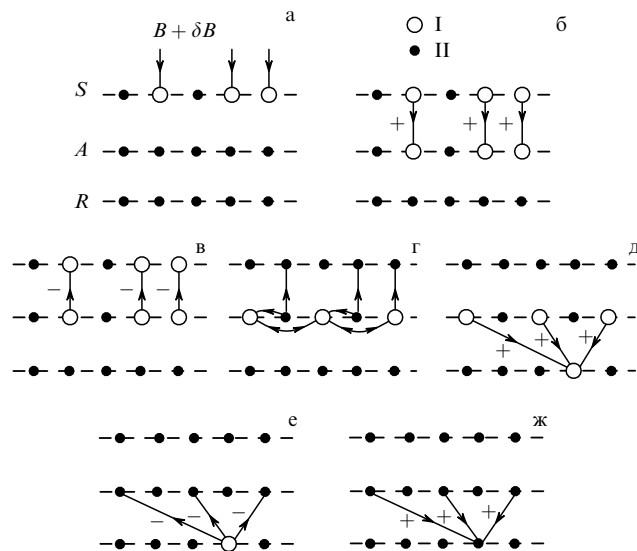


Рис. 9. Последовательные состояния системы в процессе распознавания отдельного образа; I — возбуждённые, II — невозбуждённые нейроны. Для наглядности показаны лишь связи, по которым возбуждение передавалось в предшествовавший момент времени.

учесть, что команда на выключение магнитного поля, изменение температуры (см. ниже) и обучение пластичных связей должна даваться лишь по предъявлении образа $\{V_i^{(s)}\}$ и лишь для нейронов области Ω_s .

4.6. Распознавание неоднозначного образа

Выше предполагалось, что предъявляемый стимул $\tilde{B} = B + \delta B$ достаточно близок к содержащемуся в памяти образу B , так что исходное состояние $\tilde{B}$ всегда релаксирует к конечному состоянию B , т.е. интерпретация образа является однозначной. На энергетическом языке это означает, что состояние $\tilde{B}$ попадает внутрь потенциальной ямы, минимум которой соответствует образу B , а эволюция происходит при нулевой температуре.

Рассмотрим теперь распознавание неоднозначного образа. Имеется в виду моделирование следующей ситуации: человеку показывают слово B и дают пояснение, что оно может принимать несколько значений; в памяти человека фиксируются образы $B + b_1$, $B + b_2$, ..., где b_1 , b_2 , ... — элементы данного пояснения; если теперь предъявить слово B для распознавания, то его интерпретация будет неоднозначной, приводя к одному из результатов $B + b_i$. Для моделирования достаточно предположить, что распознавание начинается при конечной температуре T_0 , превышающей потенциальные барьеры между образами $B + b_i$, а затем температура понижается и система попадает в один из соответствующих локальных минимумов. В дальнейшем будем считать, что понижение температуры от исходного значения T_0 происходит адиабатически, так что система с подавляющей вероятностью релаксирует в наиболее глубокий из локальных минимумов.

4.7. Одновременное распознавание нескольких образов

Пусть в результате предъявления стимула A система может релаксировать к одному из образов A_1 , A_2 , ..., а в результате предъявления стимула B — к одному из образов B_1 , B_2 , ... Что произойдёт при одновременном предъявлении стимулов A и B ?

В этом случае образы A и B проецируются в S -слой одновременно и распознаются как один составной многозначный образ. Пусть носители образов A и B локализованы в областях Ω_A и Ω_B ассоциативного слоя. Если области Ω_A и Ω_B не перекрываются, то образы A и B совершенно независимы. При включении связей T_{ij} между областями Ω_A и Ω_B образы A и B начинают оказывать друг на друга воздействие, эквивалентное наличию магнитного поля. Результат этого взаимодействия особенно нагляден, если конфигурации A_1 , A_2 , ... и соответственно B_1 , B_2 , ... почти вырождены. Если связи T_{ij} между областями Ω_A и Ω_B отсутствуют, то равновесные состояния системы имеют вид (A_s, B_s) и являются почти вырожденными. Включение слабых связей T_{ij} между областями Ω_A и Ω_B практически не меняет равновесных конфигураций, но изменяет взаимное расположение их уровней энергии. В результате, если при раздельном распознавании образов A и B энергетически выгодны одни конфигурации (например, A_1 и B_2 на рис. 10), то при одновременном распознавании A и B более выгодными могут оказаться другие конфигурации (например, A_2 и B_1); наглядно можно представлять, что взаимодействие меняет потенциальный рельеф, в котором релаксируют A и B (см. рис. 10). Таким образом, интерпретация многозначного образа оказывается зависящей от контекста.

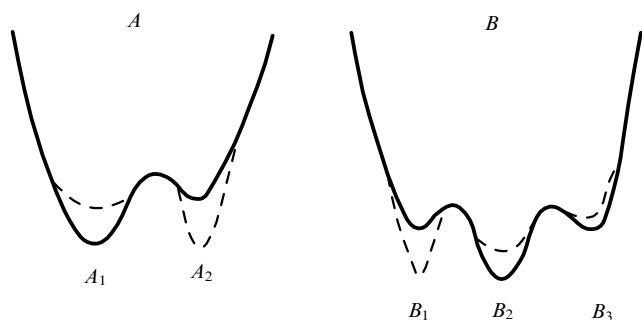


Рис. 10. Потенциальный рельеф при распознавании образов A и B поодиночке (сплошная кривая) и при одновременном распознавании (штриховая).

В разделе 3.1 предполагалось существование алгоритма для распознавания N одновременно предъявляемых многозначных образов; из изложенного ясно, что такой алгоритм естественным образом реализуется в модели Хопфилда при следующих условиях: а) распознавание начинается при конечной температуре T_0 , которая затем адиабатически понижается; б) потенциальные барьеры между значениями многозначного образа меньше T_0 ; в) потенциальные барьеры между различными многозначными образами значительно больше T_0 ; г) взаимодействие между образами является достаточно слабым. При нарушении этих условий анализ ситуации не будет столь прозрачным, но сохраняется главный качественный вывод — зависимость интерпретации от контекста.

4.8. Распознавание непрерывной последовательности образов

Выше мы предполагали, что все образы $\{V_i^{(s)}\}$ имеют локализованные носители. Ввиду конечности радиуса действия связей T_{ij} ассоциативно связанные образы должны иметь близко расположенные или перекрывающиеся носители, тогда как некоррелированным образам соответствуют носители, локализованные в удалённых участках памяти. Поэтому при поступлении в мозг непрерывной последовательности стимулов $A_1, A_2, A_3, \dots$ в ассоциативном слое происходит последовательное возбуждение областей $\Omega_1, \Omega_2, \Omega_3, \dots$, которое выглядит как "диффузия" (рис. 11). Распознавание производится при включённых связях $S \rightarrow A$ и выключенных связях $A \rightarrow S$.

Ближайшие по порядку следования образы скоррелированы, и их носители образуют скопления, тогда как удалённым по порядку следования образам вследствие ослабления корреляций соответствуют удалённые в пространстве носители. Это позволяет установить соответствие времени появления образа с положением его носителя в пространстве, что значительно упрощает работу Центра. Зафиксировав появление возбуждённых нейронов в некотором участке A -слоя, Центр повышает температуру этого участка до величины T_0 , а спустя характерное время τ_0 начинает её адиабатическое понижение, что стимулирует установление стационарного состояния, вывод в R -слой соответствующих образов и обнуление рассматриваемого участка памяти за счёт включения связей $R \rightarrow A$ (рис. 9г–ж). Тем самым траектория "диффузионного" движения (см. рис. 11) спустя некоторое время "стирается", так что в каждый момент в памяти находится лишь её конечный отрезок; тем самым

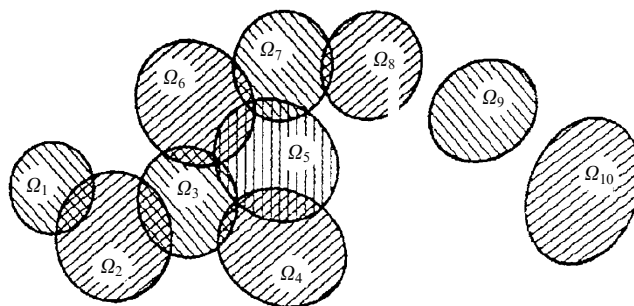


Рис. 11. При распознавании непрерывной последовательности образов $A_1, A_2, A_3, \dots$ в ассоциативном слое происходит последовательное возбуждение областей $\Omega_1, \Omega_2, \Omega_3, \dots$, которое выглядит как "диффузия".

распознавание непрерывной последовательности образов сводится к одновременному распознаванию их конечного числа (см. выше).

4.9. Временные задержки и юмористический эффект

Пусть в момент времени $t = 0$ происходит возбуждение нейронов в некотором участке памяти (рис. 9б), в момент времени τ_0 соответствующий образ выводится в сознание (рис. 9д), а в момент времени τ_1 происходит обнуление рассматриваемого участка памяти (рис. 9е). Задержка τ_0 соответствует интервалу AC , а задержка τ_1 — интервалу AB на рис. 2 в разделе 3; последнее связано с тем, что возможность переинтерпретации образа сохраняется до тех пор, пока не обнулён соответствующий участок памяти. Очевидно, $\tau_1 \geq \tau_0$; если в разделе 3.3 этот результат выводился из соображений оптимальности работы алгоритма, то теперь он обусловлен конструктивными особенностями модели.

Задержка τ_0 определяется скоростью понижения температуры (см. выше), задержка τ_1 — моментом включения из Центра $R \rightarrow A$ -связей. Оптимальный выбор параметров τ_0 и τ_1 определяется разными принципами: если параметр τ_0 обуславливает задержку от момента поступления информации в мозг до её вывода в сознание и ограничен сверху величиной $\tau_{\max}$ (см. раздел 3.3), то параметр τ_1 регулирует загрузку памяти (в отличие от общего случая, в данной модели нет специальной оперативной памяти), т.е. долю возбуждённых участков в A -слое (см. рис. 11). Эта доля должна быть не слишком мала, чтобы полнее использовать оперативные возможности мозга, и не слишком велика, во избежание интерференции образов, поступивших в разное время.

Пусть в момент $t = 0$ в память поступил образ A ; эволюция в соответствующем потенциальном рельефе (рис. 10, сплошная кривая) приводит в момент $t = \tau_0$ к стабилизации нейронов памяти в конфигурации A_1 и возбуждению соответствующего R -нейрона. Пусть в интервале между τ_0 и τ_1 в память поступил образ B и изменил потенциальный рельеф для A (см. рис. 10, штриховые линии). Если температура в рассматриваемый момент достаточна для преодоления барьера⁷, то система начинает релаксировать к конфигурации A_2 (фактически уход из состояния A_1 возможен и при $T = 0$, так как

⁷ Для этого пороги возбуждения R -нейронов должны подбираться так, чтобы возбуждение R -нейрона происходило несколько раньше, чем достигается полное равновесие и температура понижается до нуля.

соответствующий A_1 минимум может исчезнуть). Такое нарушение стационарности некоторого участка памяти после вывода в R -слой соответствующего ему образа и является признаком юмористического эффекта (см. раздел 3.3). При этом образ A_1 (в общем случае — версия из нескольких образов) осознаётся как "ложный" и должен быть немедленно выброшен из сознания; однако это не может быть сделано по обычной схеме (рис. 9д, ж) из-за необходимости получения новой стационарной конфигурации A_2 .

5. О механизме смеха

"Аварийный" сброс ложной версии осуществляется включением связей между R -слоем и моторной корой (см. рис. 7); возбуждение R -нейронов передаётся мотонейронам, вызывая сокращение определённых мышц, т.е. смех.

Нетрудно заметить сходство со старой идеей Спенсера [11], согласно которой при юмористическом эффекте происходит высвобождение из мыслительного процесса нервной энергии, которая направляется на мышечное реагирование. Эта идея поддерживалась Дарвином [8] и Фрейдом [7], но критиковалась более поздними исследователями [46] ввиду трудности введения понятия "нервной энергии". Действительно, понятие энергии в модели Хопфилда можно ввести лишь при условии $T_{ij} = T_{ji}$, которое не выполняется в реальных нейронных сетях. Однако, согласно численным экспериментам [1], введение асимметрии связей между нейронами не приводит к качественному изменению динамики системы. Поэтому качественная картина согласуется с гипотезой Спенсера и в целом действительно выглядит так, как если бы энергия возбуждения нейронов сбрасывалась в моторную кору.

Освобождение нервной энергии при юмористическом эффекте обосновывалось Спенсером с помощью концепции "нисходящей несообразности" — перехода от высокого стиля к низкому, т.е. от состояния с богатыми ассоциациями к состоянию с бедными ассоциациями. Такая трактовка юмористического эффекта является заведомо ограниченной и не может претендовать на универсальность. В предлагаемой схеме освобождение нервной энергии (в указанном выше условном смысле) связано с необходимостью избавиться от выведенной в сознание ложной версии, что требует "обнуления" некоторого участка нейронной сети. Это обстоятельство имеет общий характер и не связано со спецификой модели, использованной в разделе 4.

С физической точки зрения "обнуление" нейронной сети означает, что все нейроны, которые были возбуждены в результате записи некоторой информации, должны быть переведены в невозбуждённое состояние. На языке спиновой модели такое обнуление может быть осуществлено путём включения сильного "магнитного поля", которое делает энергетически невыгодным состояние "спин вверх" ($V_i = 1$); тогда за некоторое характерное время, определяемое эволюцией (5), система расслабляется к состоянию, в котором все спины будут направлены вниз. Произвести более быстрое стирание информации можно путём присоединения к обнуляемому участку нейронной сети "резервуара" — другого участка сети, значительно большего объёма: тогда после включения магнитного поля избыточная энергия спинов сбрасывается в резервуар. Роль резервуара могут выполнять участки нейронной сети, не связанные с функцией мышле-

ния (чтобы не возбуждать паразитных мыслительных процессов), в частности мышечные центры. В таком случае процесс стирания из сознания неправильной версии вызывает сокращение определённых мышц, т.е. смех.

Поскольку смех трактуется как безусловный рефлекс на юмористический эффект, требуют объяснения известные случаи "вытеснения" смеха побочными эмоциями. Смех может вытесняться эмоциями негодования (человеку пуританских убеждений рассказали неприличный анекдот), страха (куст внезапно оказался медведем), сострадания (идущий впереди вас человек поскользнулся на арбузной корке и сильно ушибся), стыда (на арбузной корке поскользнулись вы) и т.д. В рамках гипотезы Спенсера [11] все эти случаи объясняются тем, что освобождаемая при юмористическом эффекте психическая энергия направляется не в мышечные центры, а в другие участки нервной системы и расходуется на образование побочной эмоции (к R -слою подключается не моторная кора, а лимбическая система). Гипотезой Спенсера объясняется также известный факт, что шутка производит наибольший эффект, если она выражена предельно лаконично [7]: это связано с тем, что лаконичность уменьшает вероятность образования побочных ассоциаций, могущих поглотить психическую энергию.

Отношение $K = E_{\text{musc}}/E_{\text{nerv}}$ мышечной энергии E_{musc} к возбуждающей их нервной энергии E_{nerv} различно для разных групп мышц. Поэтому, сбрасывая возбуждение нейронов в различные мышечные центры, человек может регулировать уровень мышечной реакции: с этим связаны её зависимость от настроения, психологической установки, наличия смеющейся аудитории [47] и т.д.

6. Развитие чувства юмора в процессе эволюции

Эволюцию чувства юмора можно, по-видимому, представлять следующим образом. На первом этапе эволюции чувство юмора отсутствует, так как оно возникает лишь при достаточно сложном алгоритме обработки информации — например, в модели раздела 3 оно существует, если число сохраняемых траекторий $M > 1$. После того как достаточная сложность алгоритма достигнута и стал возможен характерный сбой, описанный в разделе 3.3, начинается второй этап эволюции, на котором чувство юмора существует, однако смех как характерная мышечная реакция отсутствует. Последнее обстоятельство связано с тем, что основная часть нейронной сети занята управлением мышечными движениями, а мыслительные процессы происходят лишь в отдельных "островках интеллекта": поэтому сброс энергии возбуждения нейронов в мышечные центры происходит случайным образом, а величина сбрасываемой энергии мала. Таким образом, на этом этапе юмористический эффект сопровождается лишь незначительными сокращениями случайных мышц и не замечается в популяции. По мере роста масштабов интеллектуальной деятельности количество энергии, сбрасываемой при юмористическом эффекте, растёт и начинает приводить к заметным сокращениям мышц: это вызывает необходимость упорядочения процесса сброса энергии с целью по возможности исключить движения конечностей (чтобы животное при смехе не теряло устойчивости). В результате сбрасываемая энергия начинает направляться в конкретные мышечные центры и возникает смех как

характерная мышечная реакция на юмористический эффект — начинается третий этап эволюции, на котором юмор осознаётся в популяции и начинает играть социальную роль.

Некоторые психологи (в частности, Фрейд [7]) придают большое значение удовольствию, получаемому от смеха, и склонны считать его главной причиной существования чувства юмора (человек открывает возможность извлечения удовольствия от психического процесса и начинает бессознательно, а затем и сознательно её эксплуатировать). С нашей точки зрения, факт получения удовольствия не является принципиальным. Действительно, пример двух сходных рефлексов — чихания и кашля — показывает, что факт получения удовольствия (при чихании) или неудовольствия (при кашле) не имеет значения для их существования, так как они вызваны биологической необходимостью очистки дыхательных путей. Аналогично, чувство юмора обусловлено биологической необходимостью ускорения вывода информации в сознание (см. раздел 3.3) и существовало бы и в том случае, если бы смех вызывал неудовольствие. Другое дело, что в последнем случае существенно изменилась бы социальная функция юмора: общество стремилось бы избавиться от него путём цензурных ограничений, преследования остроумных людей и т.д.

7. Как реализовать компьютерное "чувство юмора"?

В настоящем разделе мы попытаемся довести абстрактную модель, представленную в разделе 3, до уровня конкретного алгоритма [5, 6].

7.1. "Лингвистическая проблема"

Как ясно из раздела 3.1, для наделения компьютера "чувством юмора" требуется решение "лингвистической проблемы" — распознавания непрерывной последовательности многозначных образов. Фактически под многозначными образами подразумеваются обычные слова (см. раздел 4.6), и аналогичная проблема возникает в исследованиях по автоматическому переводу [21–28]⁸. Легко видеть, что она распадается на несколько более конкретных проблем, так что требуется:

- 1) составить список образов, имеющих хождение в популяции;
- 2) каждому слову изучаемого языка сопоставить набор ассоциированных с ним образов;
- 3) составить достаточно длинный текст для обучения, записанный непосредственно в образах;
- 4) сформулировать алгоритм обучения, сводящийся к корреляционному анализу сочетаемости образов;
- 5) на его основе выработать оценку вероятности для конечной последовательности образов.

Проблемы 1–3 в принципе тривиальны, но требуют громадного объёма достаточно квалифицированной работы; трудно представить, чтобы кто-то занимался этим

специально. Поэтому далее мы рассмотрим возможность решения этих проблем в автоматическом режиме, ориентируясь на лингвистическую информацию, собранную в электронных словарях типа ABBYY Lingvo. Проблемы 4 и 5 более сложные: они не могут быть решены на умозрительном уровне и требуют длительного экспериментирования методом "проб и ошибок". Мы предложим лишь предварительный вариант их решения, пытаясь подчеркнуть некоторые принципиальные моменты. Практический вариант алгоритма желательно формулировать с использованием имеющихся разработок для проблемы автоматического перевода [21, 22, 28].

7.2. Идеальный язык как предельный случай

Предположим, что существует язык (для определённости пусть это будет латынь), который в некотором приближении может рассматриваться как идеальный. Под идеальным мы понимаем язык, для которого слова и образы находятся во взаимно-однозначном соответствии.

Тогда проблемы 1–3 решаются тривиально. Для того чтобы составить список актуальных образов (проблема 1), достаточно переписать все латинские слова и перенумеровать их в алфавитном порядке. Если мы собираемся распознавать тексты на русском языке, то проблема 2 решается с помощью русско-латинского словаря: достаточно выписать все варианты перевода русского слова на латинский язык. В качестве текста для обучения (проблема 3) берётся произвольный литературный текст на латинском языке. Обсудим теперь проблемы 4 и 5.

Алгоритм обучения. Простейший способ обучения состоит в составлении корреляционной матрицы S_{ij} , где i и j пробегает номера образов. Перед началом обучения положим $S_{ij} \equiv 0$. Один шаг обучения будет состоять в анализе отдельного предложения, каждое из которых по определению выражает законченную мысль; поэтому все содержащиеся в нём образы (отождествляемые со словами в идеальном языке) являются взаимосвязанными. Для каждой пары слов (i, j) , входящих в данное предложение, увеличим на единицу элемент S_{ij} корреляционной матрицы,

$$\Delta S_{ij} = 1, \quad (12)$$

включая и случай $i = j$ (см. ниже). Разумеется, при таком способе обучения возникают ошибки: синтаксические связи в предложении имеют древесный характер, а существование ассоциативных связей гарантируется лишь между синтаксически связанными образами. Однако совершенно некоррелированные образы попадают в одно предложение лишь с вероятностью $\sim (n/N)^2$ (n — число слов в предложении, N — число слов в языке), что нужно сравнить с вероятностью $\sim k(n/N)$ для ассоциативно связанных образов (k — коэффициент корреляции). Ясно, что такая процедура позволяет эффективно обнаруживать корреляции между образами в широком диапазоне значений k ($1 \gtrsim k \gtrsim 10^{-3} - 10^{-4}$).

Альтернативы такому способу обучения практически нет, так как синтаксический анализ можно проводить в автоматическом режиме лишь с большим процентом ошибок [22]. С другой стороны, основывать обучение только на синтаксических связях было бы тоже не очень правильно: предложения часто содержат описание повторяющихся ситуаций (например, *Пастух гонит стадо*), в которых все образы являются ассоциативно связанными. В качестве последнего аргумента заметим, что

⁸ Несмотря на аналогию с проблемой машинного перевода, имеются некоторые различия в постановке задачи, которые в основном и являются предметом дальнейшего обсуждения. Из трёх стадий машинного перевода (анализ исходного текста, собственно перевод и генерация нового текста) для нас актуальна лишь первая, но доведённая до максимально возможного уровня. В действующих программах машинного перевода эта стадия не является достаточно разработанной (см. рис. 6.1 в [22]).

при обучении человека синтаксический анализ, по-видимому, не играет большой роли: дети и малообразованные люди владеют разговорным языком вполне сносно, хотя и не имеют представления о синтаксисе.

Из приведённых оценок ясно, что обучение желательнее проводить на коротких предложениях — с уменьшением n ошибка обучения уменьшается. Отсечение же длинных предложений не требует существенной затраты времени и ресурсов.

Вероятность траектории. После того как корреляционная матрица S_{ij} сформирована в результате обучения на достаточно длинном тексте, вероятность некоторой конечной последовательности образов можно определить как⁹

$$p = \sum_{i,j (i \neq j)} S_{ij}, \quad (13)$$

где i и j пробегает номера образов, задействованных в данной последовательности. В основу определения вероятности положен чисто аналоговый принцип, так что наиболее вероятными оказываются комбинации образов, которые часто встречаются в обучающем тексте. В определении (13) мы исключили члены с $i = j$, так как интересующая нас вероятность должна характеризовать связность траектории.

На первый взгляд, аналоговый характер алгоритма требует, чтобы условие $i \neq j$ было введено и в правило обучения (12). В действительности это не так, поскольку обучение и распознавание производятся в разных условиях: для обучения мы используем замкнутые предложения, тогда как распознаванию подвергается произвольный отрезок текста, независимо от границ предложения. Если в (12) наложить условие $i \neq j$, то самокорреляция образов окажется почти нулевой: повторное употребление слова в коротком предложении считается стилистической ошибкой¹⁰ и практически не встречается в литературном тексте. Из общих же принципов ясно, что корреляция образа самого с собой является максимальной, что и обеспечивается включением случая $i = j$ в правило обучения (12). В процессе распознавания такая самокорреляция образов имеет существенное значение: связный текст обычно характеризуется наличием "главного героя", образ которого присутствует почти в каждом предложении. Поэтому наличие одинаковых образов в соседних предложениях является вполне типичным и их самокорреляция существенна для адекватной оценки связности текста.

7.3. Реальный язык: никто не хотел "как хуже"

Разумеется, любой реальный язык очень далёк от идеального: с одной стороны, каждое слово ассоциируется со множеством образов, с другой стороны, один образ может выражаться разными словами. Возникает впечатление, что вмешался какой-то "злой дух" и специально испортил все существующие языки¹¹. На самом деле никто не хотел "как хуже" и никто не ставил целью специально запутать ситуацию: неоднозначность актуальных языков является естественным следствием их развития.

⁹ Нормировка вероятности оказывается произвольной, но это несущественно для сравнения траекторий.

¹⁰ Если это не артикль или служебное слово.

¹¹ Легенда о Вавилонской башне возникла, по-видимому, под влиянием этого впечатления.

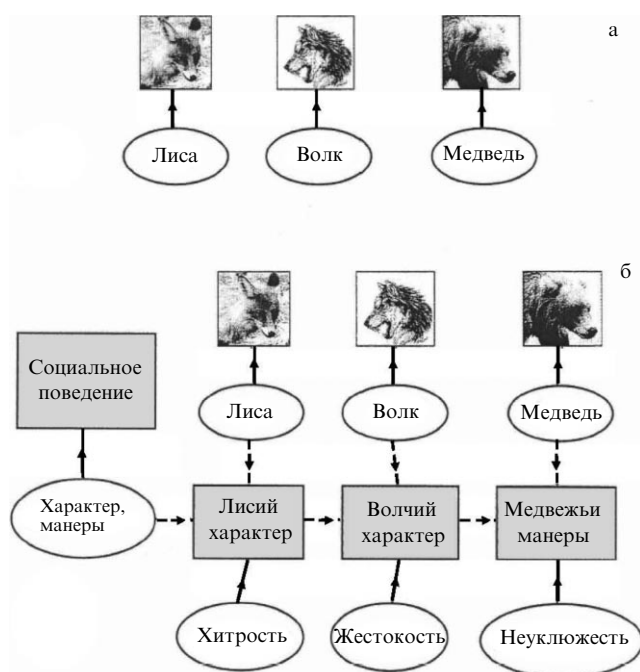


Рис. 12. (а) Первый этап развития языка: слова и образы находятся во взаимно однозначном соответствии. (б) Следующий этап развития языка: показан основной механизм возникновения его неоднозначности. Главные значения слов показаны сплошными стрелками, второстепенные значения — штриховыми.

Первые слова древнего человека формировались по принципу "что вижу, то пою"¹²: например, слова *лиса*, *волк*, *медведь* возникли из восклицаний охотников, предупреждающих соплеменников о появлении соответствующего животного. Естественно, эти слова чётко ассоциировались с указанными образами и не сопровождались никакой неоднозначностью (рис. 12а). Усложнение взаимоотношений между людьми сделало актуальными вопросы социального поведения и привело к появлению слов типа *характер*, *манеры*. Комбинирование этих слов с уже существующими привело к возникновению составных понятий: *лисий характер*, *волчий характер*, *медвежьи манеры*; эти понятия оказались полезными, и для их обозначения появились специальные слова — *хитрость*, *жестокость*, *неуклюжесть* и т.д. (рис. 12б). Нетрудно видеть, что язык адекватно отреагировал на изменение ситуации: возникли новые образы — появились новые слова; при этом количество слов осталось равным количеству образов, актуальных для популяции (рис. 12б). Однако уже на этом этапе возникает запутанность языка: с одной стороны, появилась синонимическая неоднозначность (можно сказать *хитрость*, а можно — *лисий манеры*), с другой — старые слова приобрели новые значения (слово *лиса* стало обозначать не только "рыжую с хвостом", но и хитрую Марью Ивановну). Мы видим, что запутанность языка является неизбежным следствием его развития: новые образы сначала описываются существующими словами, а затем для них придумывают специальные названия; но ассоциативных связей со старыми словами никто не отменяет.

К счастью, запутанность языка, возникающая в результате его развития, является легко устранимой. У каждого слова можно выделить его главное значение

¹² Обычай некоторых азиатских племён.

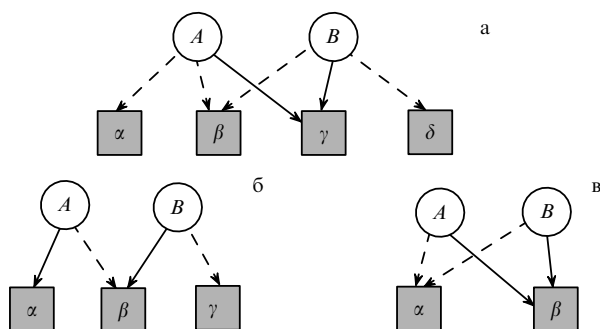


Рис. 13. (а) Слова *A* и *B* — неприводимые синонимы, их главные значения соответствуют одному образу γ . (б) Слова *A* и *B* являются синонимами по отношению к образу β , но они приводимы, так как их главные значения различны. (в) Для полноценных синонимов совпадают как главные, так и второстепенные значения.

(сплошные стрелки на рис. 12б) и второстепенные значения (штриховые стрелки). Если сопоставлять каждому слову образ, ассоциированный с его главным значением, то взаимно однозначное соответствие между словами и образами восстанавливается (рис. 12б).

К сожалению, неоднозначность языка возникает и по другим причинам, которые с точки зрения самого языка являются внешними. Так, если в двух провинциях одно и то же понятие выражалось разными словами, то при объединении этих провинций в одно государство приходится признавать допустимыми оба слова. Это приводит к возникновению неприводимых синонимов (рис. 13а), под которыми мы понимаем слова с одинаковым главным значением, в отличие от приводимых (рис. 13б), для которых главные значения различны. К такому же эффекту приводит социальное расслоение общества: так, сексуально окрашенные образы у бедных слоёв населения ассоциировались с нецензурными словами; с другой стороны, для них существуют нейтральные наименования (возникшие в аристократической среде) и медицинские термины латинского происхождения. Тот же механизм может действовать и в обратную сторону: если одни и те же слова в разных сообществах использовались для обозначения принципиально различных объектов (например, слово *коса* в русском языке), то возникают омонимы, т.е. слова с несколькими главными значениями.

Неприводимые синонимы и омонимы фактически являются дефектами языка; они возникли по случайным причинам и их существование не имеет логических оснований.

7.4. Реальный язык в качестве идеального

Проведённый в разделе 7.3 анализ позволяет понять, какие коррекции нужно внести в алгоритм, если в качестве идеального (латынь в разделе 7.2) мы попытаемся использовать реальный язык (например, английский).

Чтобы получить список образов, достаточно (в первом приближении) переписать все английские слова и ассоциировать их с главными значениями (т.е. слово считается символическим обозначением образа, который соответствует его главному значению)¹³. В первом приближении сохраняется и весь алгоритм раздела 7.2: набор образов,

ассоциируемых с русским словом, получается с помощью русско-английского словаря, для обучения используются английские тексты и т.д. Главная неточность, которая при этом совершается, состоит в том, что корреляционные связи реальных образов подменяются корреляцией английских слов. Однако с этой неточностью, по-видимому, нужно смириться, так как человек поступает точно так же. В человеческой практике не принято составлять списки образов, тогда как списки слов (словари) легко доступны. Не существует и достаточно длинных текстов, записанных в образах, но доступны длинные тексты, записанные в словах (книги). Поэтому при обучении человека неизбежно происходит подмена образов на обозначающие их слова. Сравнительно небольшая погрешность такого обучения связана со следующим:

а) Для большинства слов их "главные" значения являются действительно главными, т.е. слово с подавляющей вероятностью употребляется в этом значении.

б) Второстепенные значения обычно логически связаны с главными (рис. 12б), поэтому корреляции всего конгломерата значений с другими образами не сильно отличаются от корреляций для главного значения.

в) В человеческой практике принято проводить обучение на некотором стандартном наборе "классических" текстов. При этом подразумевается, что "классики" выражают свои мысли более чётко, чем большинство населения, так что представляемый ими образ, как правило, выражается словом, имеющим его в качестве главного значения. Так или иначе, на одинаковых текстах люди обучаются примерно одинаково (даже если и не совсем правильно), что способствует их взаимопониманию.

Теперь посмотрим, какие поправки на неидеальность языка можно реально сделать.

Неприводимые синонимы. Наличие неприводимых синонимов приводит к тому, что в составленном нами списке образов некоторые образы будут повторяться по нескольку раз. Для упорядочения ситуации удобно исходить из концепции, что полноценных синонимов (рис. 13в) практически не бывает: даже если в какой-то момент слова *A* и *B* были эквивалентны (для них совпадали как главные, так и второстепенные значения), то с течением времени они расходятся, по-разному обрастая новыми значениями (рис. 13а), и начинают употребляться в различных контекстах¹⁴. С этой точки зрения, неприводимые синонимы отмечают некоторые оттенки основного значения, т.е. соответствуют близким, но не совпадающим образам.

Сочетаемость этих слегка различных образов с другими образами адекватно описывается корреляционной матрицей, получаемой по правилу обучения (12) (поскольку фактически корреляции изучаются между словами). Проблемы связаны лишь с корреляцией между самими синонимами и аналогичны тем, что возникают в связи с самокорреляцией образов (см. раздел 7.2). Наличие двух близких образов в коротком предложении столь же нежелательно, как и повторное употребление одного и того же слова. Поэтому описанная в разделе 7.2 процедура обучения приведёт к практическому отсутствию корреляций между близкими образами, тогда как фактически эти корреляции очень сильны. Однако этот дефект можно устранить следующим образом.

¹³ Полезно сопоставить списки образов, получаемых при использовании разных языков. Образы, которые в одном языке выражаются лишь группой слов, в другом языке могут обозначаться отдельными словами.

¹⁴ Для ситуации на рис. 13а: если в некотором контексте допустимо появление образов α и γ , то образ γ должен обозначаться словом *B*; употребление слова *A* приводит к реальным недоразумениям.

В качестве модели неприводимых синонимов можно принять, что появляющийся образ с вероятностью p_A обозначается словом A , с вероятностью p_B — словом B , с вероятностью p_C — словом C и т.д.¹⁵ Повторное появление одного образа в коротком предложении является маловероятным, и типичная ситуация соответствует наличию двух одинаковых образов в соседних предложениях. Вероятности конфигураций AA , AB , ... равны соответственно p_A^2 , $p_A p_B$, ..., так что блок корреляционной матрицы, соответствующий синонимам A , B , C ... должен иметь структуру

$$\begin{pmatrix} p_A^2 & p_A p_B & p_A p_C & \dots \\ p_B p_A & p_B^2 & p_B p_C & \dots \\ p_C p_A & p_C p_B & p_C^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}, \quad (14)$$

что можно переписать в виде

$$\begin{pmatrix} S_{AA} & \sqrt{S_{AA}S_{BB}} & \sqrt{S_{AA}S_{CC}} & \dots \\ \sqrt{S_{BB}S_{AA}} & S_{BB} & \sqrt{S_{BB}S_{CC}} & \dots \\ \sqrt{S_{CC}S_{AA}} & \sqrt{S_{CC}S_{BB}} & S_{CC} & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}. \quad (15)$$

При обучении согласно алгоритму (12) диагональные элементы S_{AA} , S_{BB} , S_{CC} ... определяются правильно, тогда как недиагональные оказываются практически нулевыми; но их можно поправить искусственно, в соответствии с матрицей (15). Таким образом, "обучение синонимам" производится специально, как это имеет место и в человеческой практике.

Омонимы. С лингвистической точки зрения, омонимы считаются разными словами, что прямо отмечается в словарях (как правило, римскими цифрами, например, *коса I*, *коса II* и т.д.). Поэтому при составлении списка образов с омонимами не связано никаких проблем и они фиксируются как разные образы. Однако на письме и в речи омонимы не различаются, что создаёт проблемы при обучении.

Применительно к рассматриваемой ситуации это означает, что "обучение омонимам" должно производиться вручную: встретив в обучающем тексте один из зарегистрированных омонимов, компьютер должен запрашивать оператора, который из них подразумевается в данном предложении. Однако такой ручной этап может быть не очень продолжительным: после того как для корреляционной матрицы получена минимальная статистика, идентификацию омонимов можно доверить компьютеру: они, как правило, употребляются в существенно различном контексте и их ассоциативные связи резко различаются.

Отсутствие образов. В составленном нами списке образов некоторые образы могут просто отсутствовать, если для их обозначения не придумано специальных слов (как правило, это относится к новым или мало употребительным образам). Последнее означает, что образ ещё не вполне прижился в популяции, и многим людям он попросту неизвестен; поэтому незнание его можно простить и компьютеру.

Разумеется, такие образы можно внести в список вручную, сопоставляя его с некоторой группой слов. Но тогда вручную придётся проводить и обучение: компью-

тер должен запрашивать, соответствует ли найденная им группа слов указанному образу или эти слова попали в одно предложение случайно.

7.5. Итоги проведённого анализа

Как ясно из представленного анализа, словарные системы типа АБВУ Lingvo содержат практически всю необходимую информацию для решения "лингвистической проблемы" на первичном уровне обработки; существенно, что эта информация доступна в удобной электронной форме. Поэтому почти вся рутинная работа для решения проблем 1–5 (см. раздел 7.1) может быть проведена в автоматическом режиме без непосредственного участия человека. В имеющихся электронных словарях чётко различаются главное значение слова (оно фиксируется как "первое значение") и его второстепенные значения, а также даётся список его синонимов. В ручном режиме требуется лишь обработка омонимов и неприводимых синонимов. Словарные системы содержат также большое число фразеологических сочетаний, многие из которых соответствуют оригинальным образам; их обработка в принципе возможна, но также требует ручной работы.

В разделе 7.2 предложен аналоговый алгоритм обучения и распознавания (проблемы 4 и 5), который должен рассматриваться как предварительный и доводиться до реалистичного уровня экспериментальным путём, с использованием существующих разработок для машинного перевода [21, 22, 28]. Можно надеяться, что проведённый анализ даёт достаточный материал для начала таких экспериментов.

8. Так может ли компьютер смеяться?

Выше мы продемонстрировали принципиальную возможность наделяния компьютера "чувством юмора". Но реально ли создать действующую компьютерную программу, которая бы "смеялась" в тех же случаях, что и человек? На наш взгляд, это вполне реально, если ограничиться шутками, основанными на переключении значений отдельного слова, т.е. случаями, когда юмористический эффект возникает на первичном уровне обработки информации. Предварительный вариант конкретного алгоритма представлен в разделе 7. Следует, однако, учитывать, что такая программа будет иметь примерно такой же уровень сложности, как средняя программа для автоматического перевода [21–28], и её создание потребует работы целой группы программистов в течение нескольких лет. Заметим, однако, что такая работа представляет собой естественный новый этап в исследованиях по машинному переводу: как правило, в существующих программах в случае неоднозначной интерпретации фразы компьютер выдаёт несколько вариантов на усмотрение оператора [21], тогда как снятие этой неоднозначности имеет первостепенное значение для регистрации юмористического эффекта. Поэтому разработку компьютера с "чувством юмора" естественно проводить на основе имеющихся программ по машинному переводу.

Обучение компьютера реагированию на более сложные образцы юмора, в которых юмористический эффект возникает на высших уровнях обработки информации, в обозримом будущем представляется трудно осуществимым: это требует более детального исследования образов, возникающих на высших уровнях. Многие из этих

¹⁵ Такая модель исходит из того, что различия между близкими образами носят символический характер, и в человеческой практике на них не обращают внимания.

образов не выражаются отдельным словом, и соответствующий анализ не может проводиться в автоматическом режиме — это требует серьезной работы психологов и лингвистов.

При обсуждении реализации "чувства юмора" в нейронных сетях (см. раздел 4) мы концентрировались на организации работы нейронной сети в реальном времени и пытались сформулировать общие принципы, позволяющие зарегистрировать юмористический эффект. Практические алгоритмы обучения нейронных сетей в настоящее время бурно развиваются, наполняя конкретным содержанием абстрактный алгоритм Хопфилда. В связи с этим мы не обсуждали реализации в нейронных сетях алгоритма решения "лингвистической проблемы", предложенного в разделе 7. На наш взгляд, это должно проводиться соответствующими специалистами на основе современных разработок.

Список литературы

- Hopfield J J *Proc. Natl. Acad. Sci. USA* **79** 2554 (1982)
- Суслов И М *Биофизика* **37** 318 (1992); Suslov I M *Biophysics* **37** 242 (1992); arXiv:0711.2058
- Суслов И М *Биофизика* **37** 325 (1992); Suslov I M *Biophysics* **37** 249 (1992); arXiv:0711.2061
- Суслов И М *Компьютерная хроника* (1) 1 (1994); Suslov I M *Neuro Quantology* **6** 253 (2008); arXiv:0711.2270
- Suslov I M, arXiv:0711.3197
- Suslov I M "A 'sense of humor' in computers and the 'linguistic problem'", in *Fall 2012 Symp. on Artificial Intelligence of Humor, AAAI FSS, Washington, DC, USA, November 2–4, 2012*, Invited talk
- Фрейд З *Остроумие и его отношение к бессознательному* (М.: Современные проблемы. Н.А. Столяр, 1925); Пер. с нем. яз.: Freud S *Der Witz und seine Beziehung zum Unbewussten* (Leipzig: Franz Deuticke, 1905); Пер. на англ. яз.: Freud S *Jokes and their Relation to the Unconscious* (London: Routledge and Kegan Paul, 1960)
- Дарвин Ч *Сочинения* Т. 5 (М.: Изд-во АН СССР, 1953) с. 773–775; Пер. с англ. яз.: Darwin Ch *The Expression of the Emotions in Man and Animals* (London: John Murray, 1872) p. 132–136
- Panksepp J *Behavioural Brain Research* **182** 231 (2007)
- Розенблатт Ф *Принципы нейродинамики. Перцептроны и теория механизмов мозга* (М.: Мир, 1965); Пер. с англ. яз.: Rosenblatt F *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms* (Washington, DC: Spartan Books, 1962)
- Спенсер Г *Физиология смеха* (СПб.: Тип. А.С. Суворина, 1881); Пер. с англ. яз.: Spencer H "The physiology of laughter" *Macmillan's Mag.* **1** (5) 395 (1860)
- Veale T *Humor* **17** 419 (2004)
- Veale T, Feysaerts K, Brône G *Humor* **19** 305 (2006)
- Demjén Z J *Pragmatics* **133** 93 (2018)
- Attardo S *The Linguistics of Humor: An Introduction* (Oxford: Oxford Univ. Press, 2020)
- Tschacher W, Haken H *Entropy* **25** 341 (2023)
- Ide N, Véronis J "Word sense disambiguation: The state of the art" *Comput. Linguistics* **24** (1) 1 (1998)
- Gorfein D S (Ed.) *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity* (Washington, DC: American Psychological Association, 2001)
- Navigli R *ACM Comput. Surv.* **41** (2) 10 (2009)
- Di Marco A, Navigli R *Comput. Linguistics* **39** 709 (2013)
- Кулагина О С, Мельчук И А (Ред.) *Автоматический перевод* (М.: Прогресс, 1971)
- Hutchins W J, Somers H L *An Introduction to Machine Translation* (London: Academic Press, 1992)
- Nakov P, Ng H T J *Artificial Intelligence Res.* **44** 179 (2012)
- Katsnelson A *Nature* **609** 208 (2022)
- Piccoli V *Translat. Interpret. Studies* **17** (1) 42 (2022)
- Weber S, Mehendru N *Business Politics* **24** (1) 96 (2022)
- Naveen P, Trojovský P *iScience* **27** 110878 (2024) DOI:10.1016/j.isci.2024.110878
- Al Rousan R, Jaradat R, Malkawi M *Cogent Social Sci.* **11** 2472916 (2025) DOI:10.1080/23311886.2025.2472916
- Wiseman R *Laughlab: The Scientific Search for the World's Funniest Joke* (London: Random House, 2002)
- Binsted K, Ritchie G, in *Intern. Conf. on Humor and Laughter, Luxembourg, September 30–October 3, 1993*
- Takizawa O et al., in *Humor. Proc. of the 12th Twente Workshop on Language Technology, TWLT 12, Enschede, 14–15 May 1996* (Eds A Nijholt, G Hulstijn) (Enschede: Univ. of Twente, 1996) p. 155
- Yokogawa T, in *2002 IEEE World Congress on Computational Intelligence. Proc. of the 2002 IEEE Intern. Conf. on Fuzzy Systems: FUZZ 2002, Honolulu, HI, USA, 2002* (Piscataway, NJ: IEEE, 2002) p. 1114, DOI:10.1109/FUZZ.2002.1006660
- Taylor J M, Mazlack L J, in *Proc. of the 26th Annual Conf. of the Cognitive Science Society, Stresa, Italy, 2004*, p. 2166
- Mihalcea R, Strapparava C *Comput. Intelligence* **22** (2) 126 (2006)
- Taylor J M, Mazlack L J, in *Proc. of the 22nd AAAI Conf. on Artificial Intelligence, AAAI-07, Vancouver, Canada, July 22–26, 2007* (Menlo Park, CA: AAAI Press, 2007) p. 1904
- Kalloniatis A, Adamidis P *Artif. Intell. Rev.* **58** 43 (2025)
- Computational humor. Wikipedia, https://en.wikipedia.org/wiki/Computational_humor
- Theories of humor. Wikipedia, https://en.wikipedia.org/wiki/Theories_of_humor
- Paulos J A *Mathematics and Humor* (Chicago, IL: Univ. of Chicago Press, 1980)
- McGhee P E, Goldstein J H (Eds) *Handbook of Humor Research* Vol. 1 (New York: Springer, 1983)
- Лук А Н *О чувстве юмора и остроумии* (М.: Искусство, 1968)
- Nash W *The Language of Humor* (London: Longman, 1985)
- Raskin V *Semantic Mechanisms of Humor* (Studies in Linguistics and Philosophy, Vol. 24) (Dordrecht: D. Reidel, 1984) DOI:10.1007/978-94-009-6472-3
- Beattie J *Essays* (Edinburgh: Printed for William Creech, 1776)
- Стивенс Ч, в сб. *Мозг* (Пер. с англ. под ред. П В Симонова) (М.: Мир, 1984); Пер. с англ. яз.: Stevens C F *Sci. Am.* **241** (3) 54 (1979)
- Berlyne D E, in *The Psychology of Humor* (Eds J H Goldstein, P E McGhee) (New York: Academic Press, 1972) p. 43, DOI:10.1016/C2009-0-22105-2
- Leventhal H, Cupchik G J *Commun.* **26** (3) 190 (1976)
- Литературная газета* № 20 (1990)
- Твен М "Публичные чтения", в кн. *Собрание сочинений* Т. 8 (М.: Правда, 1980); Пер. с англ. яз.: Twain M, in *Mark Twain's Autobiography* (Ed. A B Paine) (New York: Harper and Brothers Publ., 1924)

Can computers laugh?

I.M. Suslov

Kapitza Institute for Physical Problems, Russian Academy of Sciences, ul. Kosygina 2, 119334 Moscow, Russian Federation
E-mail: suslov@kapitza.ras.ru

We propose a computer model of 'a sense of humor'. The humorous effect is interpreted as a characteristic failure in information processing, caused by the need to quickly discard a false version of the information presented to the mind. The biological function of a sense of humor is to accelerate the processing of information into the mind and to more fully utilize the brain's resources. We analyzed the feasibility of implementing the proposed algorithm in neural networks, which requires organizing their operation in real time. The mechanism of laughter and the development of a sense of humor during evolution are discussed. We propose a specific algorithm for addressing the 'linguistic problem'—namely, the recognition of a continuous sequence of ambiguous patterns, which is essential for the practical development of a computer endowed with a 'sense of humor'.

Keywords: humor, laughter, language models, Hopfield network

PACS numbers: 07.05.Mh, **87.19**, –j, 89.20.Ff

Bibliography — 49 references

Uspekhi Fizicheskikh Nauk **196** (10) 1098–1113 (2026)

DOI: <https://doi.org/10.3367/UFNr.2026.08.040190>

Received 17 November 2025

Physics – *Uspekhi* **69** (10) (2026)

DOI: <https://doi.org/10.3367/UFNe.2026.08.040190>