Ординальная агрегация наборов векторных представлений слов

А. М. Колосов*

В работе исследуется задача слияния наборов векторных представлений слов от различных предобученных моделей. Поскольку прямое объединение векторных пространств сопряжено с вычислительными трудностями и проблемой локальных минимумов, предлагается переход к ординальным методам агрегации. Доказывается отсутствие циклов в индивидуальных ординальных матрицах попарных сравнений расстояний, что гарантирует эквивалентность алгоритмов Кемени и Коупленда в этом случае. Для агрегированных матриц, где возможен парадокс Кондорсе, алгоритм Коупленда выступает эффективной эвристикой. Также рассматривается метод Борда, не требующий построения турнирных графов. Проведены эксперименты на 10 моделях и 3 датасетах (WordSim-353, MEN, SimLex-999). Показано, что ординальные методы позволяют находить оптимальные комбинации моделей (тройки, четвёрки и пятёрки), превосходящие индивидуальные модели по корреляции Спирмена, при этом разница между методами Борда и Кемени/Коупленда минимальна.

Ключевые слова: векторные представления слов, агрегация моделей, метод Борда, метод Кемени, алгоритм Коупленда, турнирные графы, ординальная матрица.

1. Введение

Векторные представления слов (word embeddings) играют ключевую роль в задачах обработки естественного языка. Различные архитектуры и обучающие корпуса приводят к тому, что модели улавливают разные аспекты семантики: например, модели на основе графов знаний (ConceptNet

Колосов Алексей Михайлович — научный сотрудник кафедры математической теории интеллектуальных систем механико-математического факультета Московского государственного университета имени М. В. Ломоносова, научный сотрудник лаборатории инженерии знаний института математических исследований сложных систем Московского государственного университета имени М. В. Ломоносова, e-mail: aleksei.kolosov@math.msu.ru, ORCID: 0000-0002-9474-9666.

Kolosov Alexey Mikhajlovich — Research Fellow, Department of Mathematical Theory of Intelligent Systems, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, Research Fellow, Knowledge Engineering Lab, Institute of Complex Systems Mathematical Research, Lomonosov Moscow State University.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

Numberbatch [1]) хорошо отражают строгие семантические связи, классические дистрибутивные модели (GloVe [2], Word2Vec [3], fastText [4]) улавливают статистические закономерности совместной встречаемости, модели на основе парафразных ограничений (Paragram [5]) оптимизированы для различения синонимов и антонимов, модели на основе контекста (SBERT [6], OpenAI text-embedding [7]) извлекают глубокий контекстный сигнал, а модели контрастивного обучения (SimCSE [8]) формируют представления, оптимизированные для различения семантически близких и далёких предложений.

Мотивация данной работы заключается в том, чтобы использовать готовые предобученные модели без необходимости обучать новые модели с нуля. Возникает задача: как эффективно объединить информацию из нескольких наборов векторных представлений, чтобы получить единую меру семантической близости, превосходящую каждую модель по отдельности? Прямое объединение (конкатенация или линейные преобразования) векторных пространств различной размерности и природы часто оказывается неэффективным. В данной статье предлагается подход, основанный на извлечении и объединении ординальной (порядковой) информации из различных векторных пространств.

2. Ординальная матрица для пар пар слов

Рассматривается множество слов W и множество всех пар слов $P = W \times W$. Каждая модель задаёт функцию расстояния (или сходства, например, косинусного) $d(p)$ для $p \in P$. Вместо того чтобы работать с абсолютными значениями $d(p)$, осуществляется переход к относительным сравнениям.

Для каждой пары пар слов $(p_i, p_j) \in P \times P$ фиксируется знак сравнения: $p_i \succ p_j$, если пара p_i семантически ближе, чем p_j в пространстве данной модели. Таким образом, каждая модель порождает индивидуальную ординальную матрицу (турнирный граф), где вершинами являются пары слов, а рёбра указывают направление предпочтения.

Пример турнирного графа для трёх пар слов: пусть $p_1 = (\text{кошка}, \text{собака})$, $p_2 = (\text{машина}, \text{автобус})$, $p_3 = (\text{яблоко}, \text{груша})$. В таблице 1 приведён пример индивидуальной ординальной матрицы для случая, когда модель оценивает сходства как $d(p_1) = 0.9$, $d(p_2) = 0.8$, $d(p_3) = 0.7$.

Лемма 1 (Отсутствие циклов в индивидуальной матрице)**.** *В индивидуальной ординальной матрице, порождённой функцией расстояния в метрическом пространстве, отсутствуют циклы.*

Таблица 1: Пример индивидуальной ординальной матрицы

Пара слов	p_1	p_2	p_3
p_1 (кошка, собака)	–	$\succ$	$\succ$
p_2 (машина, автобус)	$\prec$	–	$\succ$
p_3 (яблоко, груша)	$\prec$	$\prec$	–

Доказательство. Предполагается противное: существует цикл $p_1 \succ p_2 \succ \cdots \succ p_k \succ p_1$. Это означает, что $d(p_1) > d(p_2) > \cdots > d(p_k) > d(p_1)$. По транзитивности строгого неравенства для действительных чисел получается $d(p_1) > d(p_1)$, то есть расстояние больше самого себя. Полученное противоречие доказывает отсутствие циклов. Важно отметить, что отсутствие циклов является следствием природы происхождения турнирного графа: каждое ребро порождается сравнением значений $d(p_i)$ и $d(p_j)$, а транзитивность строгого неравенства на действительных числах гарантирует ацикличность. Произвольный турнирный граф, не порождённый функцией расстояния, может содержать циклы. □

Для агрегированной ординальной матрицы, получаемой путём голосования большинства моделей, лемма 1 уже не выполняется. Известен парадокс Кондорсе [9]: возможна ситуация, когда большинство моделей предпочитает A перед B, большинство предпочитает B перед C, и большинство предпочитает C перед A. В таблице 2 приведён пример возникновения такого парадокса при агрегации трёх моделей.

Таблица 2: Пример возникновения парадокса Кондорсе

Модель	**1-е место**	**2-е место**	**3-е место**
Модель 1	A	B	C
Модель 2	B	C	A
Модель 3	C	A	B
Попарные победы (большинство 2:1): $A \succ B$, $B \succ C$, $C \succ A$ — цикл			

Теоретическим обоснованием возможности восстановления метрического пространства по ординальным ограничениям служит теорема Агарвала [10], утверждающая существование точного вложения в евклидово пространство при выполнении определённых условий на ординальные ограничения. К таким условиям относятся требования к размерности пространства вложения (которая должна быть достаточно большой) и

согласованности системы неравенств (отсутствие противоречивых ограничений вида $d_1 > d_2$ и $d_2 > d_1$).

3. Базовый подход: алгоритм GNMDS

Одним из способов решения задачи является восстановление векторных представлений с сохранением рангов (ординальных ограничений). Базовым подходом здесь является обобщённое неметрическое многомерное шкалирование (Generalized Non-metric Multidimensional Scaling, GNMDS) [11].

Этот алгоритм использует градиентный спуск для поиска координат векторов, которые минимизируют нарушения заданных неравенств $d(p_i) > d(p_j)$. Однако GNMDS имеет существенные недостатки:

1. Алгоритм основан на градиентном спуске и не гарантирует нахождения глобального минимума.
2. Высокая вычислительная сложность, ограничивающая масштабируемость на большие словари.

Поскольку качество векторных представлений в стандартных бенчмарках оценивается с помощью ординального показателя качества — корреляции Спирмена ρ, которая измеряет монотонную связь между предсказанными рангами пар слов и рангами экспертных оценок (то есть вычисляется как корреляция Пирсона между рангами двух переменных, принимая значения от -1 до $+1$), — восстановление полных векторов избыточно. Для решения задачи достаточно восстановить итоговые ранги пар слов.

4. Первый подход: метод Кемени и алгоритм Коупленда

Переход от ординальной матрицы непосредственно к рангам можно осуществить методами теории социального выбора. Оптимальным подходом к поиску консенсусного ранжирования по турнирному графу является метод Кемени (Kemeny-Young method) [12]. Задача Кемени состоит в поиске линейного порядка, минимизирующего число инверсий относительно рёбер турнирного графа (расстояние Кендалла).

К сожалению, задача Кемени является NP-трудной [9]. В качестве практической эвристики используется алгоритм Коупленда (Copeland's method) [13], который ранжирует вершины турнирного графа по числу попарных побед (исходящих рёбер).

Теорема 1 (Эквивалентность Кемени и Коупленда без циклов)**.** *Если в турнирном графе отсутствуют циклы (граф является транзитивным турниром), то ранги, полученные по алгоритму Коупленда, в точности совпадают с оптимальными рангами по Кемени, и оба метода эквивалентны топологической сортировке графа.*

Доказательство. Пусть турнирный граф G на n вершинах не содержит циклов. Тогда G является транзитивным турниром: для любых трёх вершин u, v, w, если $u \succ v$ и $v \succ w$, то $u \succ w$.

В транзитивном турнире существует единственная топологическая сортировка $\sigma = (v_1, v_2, \ldots, v_n)$, такая что $v_i \succ v_j$ для всех $i < j$.

Коупленд даёт σ: Вершина v_k побеждает ровно $n - k$ вершин (все v_j с $j > k$). Следовательно, число побед строго убывает: $n - 1, n - 2, \ldots, 0$. Сортировка по числу побед однозначно даёт порядок σ.

Кемени даёт σ: Порядок σ согласован со всеми $\binom{n}{2}$ рёбрами графа (ни одно ребро не направлено "против" σ), поэтому число инверсий равно нулю. Это глобальный минимум расстояния Кендалла, так как число инверсий неотрицательно. Любой другой порядок $\pi \neq \sigma$ имеет хотя бы одну инверсию (хотя бы одно ребро направлено против π, поскольку σ — единственный порядок, согласованный со всеми рёбрами транзитивного турнира).

Таким образом, оба метода дают один и тот же результат — топологическую сортировку σ. □

Из леммы 1 и теоремы 1 следует, что для индивидуальной ординальной матрицы (одной модели) алгоритм Коупленда всегда даёт оптимальное решение задачи Кемени. Для агрегированной матрицы, где циклы возможны, алгоритм Коупленда является лишь эвристикой для задачи Кемени и не гарантирует нахождения глобального оптимума, однако на практике демонстрирует высокую эффективность и вычислительную простоту ($O(N^2)$ для графа из N вершин).

5. Второй подход: метод Борда

Альтернативным подходом является метод Борда (Borda count) [14]. В отличие от Кемени и Коупленда, метод Борда работает непосредственно с рангами (позициями элементов в индивидуальных линейных порядках), а не со знаками попарных сравнений. Ординальная матрица для этого метода не строится. Каждой паре слов присваиваются баллы в зависимости от её позиции в ранжировании каждой модели: элемент на позиции k из n получает $n - k$ баллов, после чего баллы суммируются по

всем моделям. Итоговые ранги определяются сортировкой элементов по убыванию суммы баллов: элемент с наибольшей суммой получает ранг 1, следующий — ранг 2, и так далее.

Хотя метод Борда вычислительно эффективнее ($O(N \log N)$), его результаты могут отличаться от Кемени/Коупленда. В таблице 3 приведён пример, демонстрирующий это различие на четырёх элементах $\{A, B, C, D\}$.

Таблица 3: Пример различия методов Борда и Кемени/Коупленда

Модель	**1-е (3 б.)**	**2-е (2 б.)**	**3-е (1 б.)**	**4-е (0 б.)**
Модель 1	A	B	C	D
Модель 2	D	C	A	B
Модель 3	D	A	B	C
Суммы Борда: $A = 6, D = 6, B = 3, C = 3$ **Результат Борда:** ничья $A \sim D$, ничья $B \sim C$				
Попарные победы: $D \succ A$ (2:1), $A \succ B$ (3:0), $D \succ B$ (2:1) **Результат Кемени/Коупленда:** $D \succ A \succ B \succ C$ (однозначно)				

Как видно из примера, два элемента (A и D) имеют одинаковую сумму баллов Борда (одинаковые средние позиции), однако в прямом попарном сравнении D побеждает A большинством голосов (2:1). Метод Борда оперирует только абсолютными позициями элементов в ранжированиях и не учитывает результаты прямых попарных сравнений, тогда как Кемени/Коупленд использует именно эту информацию для однозначного разрешения ничьих.

6. Описание моделей и датасетов

В экспериментах использовались 10 предобученных моделей, охватывающих различные подходы к получению векторных представлений:

1. **ConceptNet Numberbatch (CN)** [1] — векторы, основанные на графе знаний ConceptNet, обогащённые семантическими отношениями. Размерность: 300.
2. **Paragram-SL999 (PG)** [5] — векторы, дообученные с использованием лингвистических ограничений для улучшения различения синонимов и антонимов. Размерность: 300.
3. **OpenAI text-embedding-3-large (3L)** [7] — современная LLM-модель эмбеддингов большой размерности (3072), извлекающая глубокий контекст.

4. **OpenAI text-embedding-3-small (3s)** [7] — компактная LLM-модель (1536 размерностей) нового поколения.
5. **OpenAI text-embedding-ada-002 (ada)** [7] — модель эмбеддингов предыдущего поколения (1536 размерностей).
6. **Sentence-BERT (SBERT)** [6] — модель на основе архитектуры BERT, дообученная для получения семантически значимых представлений предложений. Размерность: 768.
7. **GloVe-300** [2] — классические векторы, полученные факторизацией матрицы глобальной совместной встречаемости слов. Размерность: 300.
8. **fastText** [4] — векторы, учитывающие морфологию через n-граммы подслов (subwords). Размерность: 300.
9. **Word2Vec** [3] — классические векторы, обученные с использованием архитектуры Skip-gram. Размерность: 300.
10. **SimCSE (sup-simcse-bert-base-uncased)** [8] — модель контрастивного обучения на основе BERT, дообученная на парах естественного языкового вывода (NLI) для получения семантически значимых представлений предложений. Размерность: 768.

Оценка качества проводилась на трёх стандартных датасетах:

- **WordSim-353 (WS353)** [15]: состоит из 353 пар слов. Оценивает общую ассоциативную и семантическую близость, не делая строгого различия между ними (например, пара “кофе” и “чашка” может иметь высокую оценку).
- **MEN** [16]: содержит 3000 пар слов. Оценивает визуальную и семантическую ассоциативность, полученную на основе суждений людей по изображениям и текстам.
- **SimLex-999** [17]: состоит из 999 пар слов. Строго оценивает именно семантическое сходство, отличая синонимы от связанных ассоциативно слов (например, “кофе” и “чашка” получат низкую оценку, а “кофе” и “чай” — более высокую). В датасет включены только пары с совпадающей частью речи.

7. Результаты индивидуальных моделей

В таблице 4 представлены результаты индивидуальных моделей (корреляция Спирмена ρ). Первые и вторые максимумы в каждом столбце выделены полужирным и подчёркиванием соответственно.

Таблица 4: Корреляция Спирмена для индивидуальных моделей

Модель	**WS353**	**MEN**	**SimLex**	**Среднее**
ConceptNet	**0.815**	**0.871**	0.627	**0.771**
Paragram	0.720	0.778	0.685	0.728
SimCSE	0.642	0.798	**0.693**	0.711
3L	0.733	0.793	0.566	0.698
SBERT	0.744	0.788	0.536	0.689
fastText	0.697	0.803	0.441	0.647
Word2Vec*	0.694	0.782	0.442	0.639
3s	0.650	0.754	0.502	0.635
ada	0.649	0.737	0.439	0.608
GloVe-300	0.609	0.749	0.370	0.576

ConceptNet демонстрирует лучшие результаты на WS353 и MEN, тогда как SimCSE лидирует на строго семантическом SimLex (0.693), опережая Paragram (0.685). Это объясняется тем, что контрастивное обучение SimCSE эффективно различает семантически близкие и далёкие пары. Индивидуальные результаты Word2Vec получены на усечённом наборе пар слов (в модели отсутствуют векторные представления для 5 слов). Это различие в составе оцениваемых пар слов незначительно влияет на абсолютные значения корреляций и не затрагивает результаты ансамблей (в состав рассматриваемых ансамблей Word2Vec не входит).

8. Поиск комбинаций методом Борда

Используя метод Борда, был проведён перебор комбинаций из 3, 4 и 5 моделей для поиска ансамблей, максимизирующих среднюю корреляцию на трёх датасетах. Для сравнения приведены результаты трёх лучших индивидуальных моделей (ConceptNet, Paragram и SimCSE). Все 10 исследованных ансамблей представлены в таблице 5.

Анализ результатов показывает, что комбинирование моделей позволяет превзойти лучшую индивидуальную модель (ConceptNet, 0.771) на ~ 0.032 в среднем. Особо стоит отметить синергетический эффект на датасете SimLex: тройка CN + PG + SimCSE достигает 0.722, что превышает лучшую индивидуальную модель SimCSE (0.693) на +0.029. Этот ансамбль превосходит лучшую индивидуальную модель на SimLex, что демонстрирует силу ординальной агрегации для строго семантических задач. Ключевую роль играет включение SimCSE: все четыре лучших

Таблица 5: Комбинации моделей (агрегация Борда)

Конфигурация	**WS353**	**MEN**	**SimLex**	**Среднее**
ConceptNet (индивид.)	0.815	0.871	0.627	0.771
Paragram (индивид.)	0.720	0.778	0.685	0.728
SimCSE (индивид.)	0.642	0.798	0.693	0.711
CN + PG + 3L + SimCSE	0.819	**0.881**	0.709	**0.803**
CN+SBERT+PG+3L+SimCSE	0.827	0.881	0.696	0.801
CN + PG + SimCSE	0.801	0.877	**0.722**	0.800
CN + PG + SimCSE + SBERT	0.819	0.878	0.705	0.800
PG + SimCSE + SBERT	0.794	0.863	0.715	0.791
CN + SBERT + PG + 3L	**0.837**	0.867	0.665	0.790
CN + PG + 3L	0.828	0.864	0.675	0.789
CN + SBERT + PG	0.829	0.860	0.669	0.786
CN + 3s + PG	0.810	0.861	0.669	0.780
CN + PG + ada	0.818	0.857	0.646	0.774

ансамбля по среднему содержат эту модель, что подтверждает ценность контрастивного сигнала для ординальной агрегации.

9. Поиск комбинаций алгоритмом Коупленда

Аналогичный перебор комбинаций был выполнен с использованием алгоритма Коупленда для агрегации. В таблице 6 представлены все 10 комбинаций моделей. Для сравнения также приведены результаты трёх лучших индивидуальных моделей.

Результаты Коупленда демонстрируют аналогичные закономерности: четвёрка CN + PG + 3L + SimCSE лидирует по среднему (0.803), а лучший результат на SimLex делят тройки CN + PG + SimCSE и PG + SimCSE + SBERT (0.713). Алгоритм Коупленда стабильно показывает несколько более высокие результаты на ассоциативных датасетах (WS353, MEN), тогда как на строго семантическом SimLex метод Борда оказывается более устойчивым.

10. Сравнение метода Борда и алгоритма Коупленда

Для трёх лучших конфигураций (по среднему Борда) было проведено прямое сравнение метода Борда и алгоритма Коупленда (как эвристики

Таблица 6: Комбинации моделей (агрегация Коупленда)

Конфигурация	**WS353**	**MEN**	**SimLex**	**Среднее**
ConceptNet (индивид.)	0.815	0.871	0.627	0.771
Paragram (индивид.)	0.720	0.778	0.685	0.728
SimCSE (индивид.)	0.642	0.798	0.693	0.711
CN + PG + 3L + SimCSE	0.822	**0.883**	0.704	**0.803**
CN + PG + SimCSE	0.809	0.879	**0.713**	0.800
CN + PG + SimCSE + SBERT	0.822	0.880	0.700	0.800
CN+SBERT+PG+3L+SimCSE	0.829	0.882	0.689	0.800
CN + PG + 3L	0.834	0.870	0.670	0.791
CN + SBERT + PG + 3L	**0.838**	0.871	0.662	0.790
PG + SimCSE + SBERT	0.792	0.861	0.713	0.789
CN + SBERT + PG	0.830	0.867	0.663	0.786
CN + 3s + PG	0.819	0.866	0.670	0.785
CN + PG + ada	0.826	0.865	0.652	0.781

для Кемени). Результаты приведены в таблице 7. В столбце $\Delta\rho$ указано абсолютное различие, а также процентное различие относительно метода Борда.

Таблица 7: Сравнение метода Борда и алгоритма Коупленда (ρ Спирмена)

Конфигурация	**Датасет**	**Борда**	**Коупленд**	$\Delta\rho$ (%)
CN+PG+3L+SimCSE	WS353	0.819	**0.822**	+0.003 (+0.4%)
	MEN	0.881	**0.883**	+0.002 (+0.2%)
	SimLex	**0.709**	0.704	−0.005 (−0.7%)
CN+SBERT+PG+ 3L+SimCSE	WS353	0.827	**0.829**	+0.002 (+0.2%)
	MEN	0.881	**0.882**	+0.001 (+0.1%)
	SimLex	**0.696**	0.689	−0.007 (−1.0%)
CN+PG+SimCSE	WS353	0.801	**0.809**	+0.008 (+1.0%)
	MEN	0.877	**0.879**	+0.002 (+0.2%)
	SimLex	**0.722**	0.713	−0.009 (−1.2%)

Результаты показывают, что методы дают близкие, но не идентичные результаты. Доля инверсий между итоговыми ранжированиями — то есть доля пар элементов (p_i, p_j), для которых порядок в методе Борда противоположен порядку в алгоритме Коупленда — составляет 2–4%. Эта величина непосредственно связана с расстоянием Кендалла τ между двумя ранжированиями: доля инверсий равна $(1 - \tau)/2$, где τ — коэффициент ранговой корреляции Кендалла между результатами двух методов.

Таким образом, 2–4% инверсий соответствуют корреляции Кендалла $\tau \approx 0.92$–0.96 между ранжированиями Борда и Коупленда.

Алгоритм Коупленда стабильно показывает лучшие результаты на ассоциативных датасетах (WS353, MEN), так как он лучше разрешает ничьи через попарное большинство. Метод Борда оказывается более устойчивым на строго семантическом SimLex. В целом, разница между методами невелика ($\Delta\rho \leq 0.009$, или $\leq 1.2\%$), что оправдывает использование вычислительно более простого метода Борда на практике.

11. Обсуждение: стратегия вето большинством

В рассматриваемых ансамблях присутствует неявная стратегия *вето большинством*: если одна модель даёт аномальный результат для некоторой пары слов (например, размещает её на последнем месте, тогда как остальные — на первых), большинство моделей эффективно нейтрализует эту ошибку. Эта стратегия работает в обоих методах, но с разной механикой.

В **методе Борда** вето реализуется через усреднение рангов. Если в тройке моделей одна присваивает паре слов ранг N (последнее место), а две другие — ранги, близкие к 1, то сумма Борда будет определяться преимущественно двумя согласными моделями, и пара окажется в верхней части итогового ранжирования. Однако влияние аномальной модели полностью не устраняется: её вклад в сумму смещает итоговый ранг вниз, хотя и не критически.

В **алгоритме Коупленда** вето работает более жёстко. В попарном сравнении двух пар слов решение принимается большинством голосов (например, 2:1 в тройке). Аномальный голос одной модели полностью игнорируется: независимо от того, насколько сильно одна модель отклоняется от консенсуса, результат попарного сравнения определяется только знаками предпочтений большинства, а не величинами отклонений. Это делает Коупленда более устойчивым к выбросам, но одновременно он теряет информацию о *степени* предпочтения, которую сохраняет метод Борда через суммы рангов.

Стратегия вето объясняет, почему ансамбли стабильно превосходят индивидуальные модели: ошибки отдельных моделей, как правило, некоррелированы (разные модели ошибаются на разных парах слов), и большинство ветирует каждую такую ошибку.

12. Обсуждение: переход от векторов слов к векторам текстов

Интересным наблюдением является то, что модели, оптимизированные для векторизации целых текстов и предложений (SBERT, OpenAI text-embedding-3), показывают посредственные результаты на уровне отдельных слов (особенно на SimLex) по сравнению со специализированными моделями (ConceptNet, Paragram). Исключением является SimCSE, которая, благодаря контрастивному обучению, достигает лучшего индивидуального результата на SimLex (0.693), несмотря на скромные показатели на WS353 (0.642). Тем не менее включение контекстных моделей в ансамбли (например, CN+PG+3L+SimCSE) стабильно повышает итоговое качество. Это свидетельствует о том, что LLM-эмбеддинги и модели контрастивного обучения содержат уникальный сигнал, который ординальные методы успешно комбинируют с графовой семантикой.

Переход от векторов отдельных слов к векторам целых текстов является логичным продолжением развития методов агрегации. Хотя данное исследование фокусируется на парах слов, предложенные ординальные подходы (Борда и Коупленд) математически не зависят от природы объектов. Они могут быть напрямую применены для агрегации оценок семантической близости целых предложений или документов, что особенно актуально для задач информационного поиска (Information Retrieval) и систем RAG (Retrieval-Augmented Generation), где объединение ранжирований от разных моделей поиска является стандартной практикой.

13. Обсуждение: от слияния векторов к слиянию моделей

Переход к ординальным методам (Борда, Кемени) открывает путь от задачи простого математического выравнивания векторных пространств к задаче ансамблирования моделей (model fusion). Вместо того чтобы пытаться найти единое пространство, модели используются как независимые “эксперты”, голосующие за семантическую близость понятий. Этот подход не требует доступа к весам моделей или обучающим данным, не подвержен проблеме локальных минимумов (как GNMDS) и легко масштабируется на любое количество новых моделей (LLM), позволяя извлекать синергетический эффект из их различных архитектур.

Важным направлением развития является слияние моделей с целью получения единых функций расстояния, применимых для получения оценок векторных представлений для новых объектов, в частности, тек-

стов. Данная задача тесно связана с областью обучения метрик (metric learning), где целью является построение функции расстояния, сохраняющей заданные ординальные ограничения на новых данных. Если ординальная агрегация на фиксированном наборе данных позволяет найти оптимальный консенсусный порядок, то обучение мета-модели (например, нейронной сети), предсказывающей этот консенсусный порядок для новых пар текстов на основе их исходных представлений, позволит обобщить результаты агрегации. Это превращает ординальные методы из инструмента оценки в инструмент создания новых, более сильных гибридных моделей векторных представлений — фактически реализуя подход metric learning на основе ансамблевых ординальных ограничений.

14. Заключение

В данной работе исследовано применение ординальных методов для агрегации наборов векторных представлений слов. Доказано, что отсутствие циклов в индивидуальных ординальных матрицах (Лемма 1) гарантирует эквивалентность алгоритмов Кемени и Коупленда (Теорема 1), что обосновывает использование вычислительно эффективного алгоритма Коупленда в качестве точного решения задачи Кемени для индивидуальных моделей и в качестве эвристики для агрегированных данных. Экспериментальное сравнение на 10 моделях и 3 датасетах продемонстрировало, что ординальные подходы (Борда и Коупленд) позволяют находить ансамбли, превосходящие лучшие индивидуальные модели. При этом разница в качестве между вычислительно простым методом Борда и более сложным алгоритмом Коупленда минимальна (не более 1.2% по корреляции Спирмена), что делает метод Борда предпочтительным выбором для практических задач. Предложенный подход открывает перспективы для ансамблирования моделей текстовых эмбеддингов без необходимости их дорогостоящего дообучения. После получения агрегированных рангов векторные представления, согласованные с консенсусным порядком, могут быть восстановлены с помощью алгоритма GNMDS [11], который минимизирует нарушения ординальных ограничений при построении евклидова вложения. Таким образом, полный конвейер включает два этапа: (а) ординальная агрегация методом Борда или алгоритмом Коупленда для получения консенсусных рангов и (б) восстановление векторных представлений алгоритмом GNMDS [11] для использования в последующих задачах.

Финансирование

Работа выполнена в рамках государственного задания МГУ имени М. В. Ломоносова «Инженерия знаний. Разработка цифровой платформы и онтологической системы фундаментальных знаний на русском языке» (№ 124020100068-4).

Список литературы

[1] R. Speer, J. Chin, C. Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge", *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, 4444–4451. DOI: 10.1609/aaai.v31i1.11164.

[2] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532–1543. DOI: 10.3115/v1/D14-1162.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", *Advances in Neural Information Processing Systems 26 (NeurIPS)*, 2013, 3111–3119. DOI: 10.48550/arXiv.1310.4546.

[4] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, **5** (2017), 135–146. DOI: 10.1162/tacl_a_00051.

[5] J. Wieting, M. Bansal, K. Gimpel, K. Livescu, "From Paraphrase Database to Compositional Paraphrase Model and Back", *Transactions of the Association for Computational Linguistics*, **3** (2015), 345–358. DOI: 10.1162/tacl_a_00143.

[6] N. Reimers, I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, 3982–3992. DOI: 10.18653/v1/D19-1410.

[7] A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder, L. Weng, "Text and Code Embeddings by Contrastive Pre-Training", *arXiv preprint arXiv:2201.10005*, 2022. DOI: 10.48550/arXiv.2201.10005.

[8] T. Gao, X. Yao, D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.

[9] J. J. Bartholdi, C. A. Tovey, M. A. Trick, "Voting schemes for which it can be difficult to tell who won the election", *Social Choice and Welfare*, **6**:2 (1989), 157–165. DOI: 10.1007/BF00303169.

[10] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, S. Belongie, "Generalized Non-metric Multidimensional Scaling", *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007, 11–18.

[11] А. М. Колосов, А. И. Майсурадзе, “Улучшение качества векторных представлений слов за счёт использования нескольких источников представлений”, *Интеллектуальные системы. Теория и приложения*, **30**:1 (2026), 87–100.

[12] J. G. Kemeny, “Mathematics without numbers”, *Daedalus*, **88**:4 (1959), 577–591.

[13] A. H. Copeland, “A “reasonable” social welfare function”, *Seminar on Applications of Mathematics to Social Sciences, University of Michigan*, 1951.

[14] J.-C. Borda, “Mémoire sur les élections au scrutin”, *Histoire de l’Académie Royale des Sciences*, 1781, 657–665.

[15] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Sasson, G. Wolfman, E. Ruppin, “Placing Search in Context: The Concept Revisited”, *Proceedings of the 10th International Conference on World Wide Web*, 2001, 406–414. DOI: 10.1145/371920.372094.

[16] E. Bruni, N.-K. Tran, M. Baroni, “Multimodal Distributional Semantics”, *Journal of Artificial Intelligence Research*, **49** (2014), 1–47. DOI: 10.1613/jair.4135.

[17] F. Hill, R. Reichart, A. Korhonen, “SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation”, *Computational Linguistics*, **41**:4 (2015), 665–695. DOI: 10.1162/COLI_a_00237.

Статья поступила 15 мая 2026 г.

Ordinal Aggregation of Word Embedding Sets

A. M. Kolosov

This paper investigates the problem of merging word embedding sets from various pre-trained models. Since direct vector space alignment faces computational challenges and local minima issues, it is proposed to transition to ordinal aggregation methods. The absence of cycles in individual ordinal matrices of pairwise distance comparisons is proven, which guarantees the equivalence of the Kemeny and Copeland algorithms in this case. For aggregated matrices where the Condorcet paradox is possible, Copeland’s algorithm serves as an efficient heuristic. The Borda count method, which does not require building tournament graphs, is also considered. Experiments were conducted on 10 models and 3 datasets (WordSim-353, MEN, SimLex-999). It is shown that ordinal methods can identify optimal model combinations (triples, quads, and quints) that outperform individual models in Spearman correlation, with minimal difference between the Borda and Kemeny/Copeland methods.

Keywords: word embeddings, model aggregation, Borda count, Kemeny rule, Copeland algorithm, tournament graphs, ordinal matrix.

References

[1] R. Speer, J. Chin, C. Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge", *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, 4444–4451. DOI: 10.1609/aaai.v31i1.11164.

[2] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532–1543. DOI: 10.3115/v1/D14-1162.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", *Advances in Neural Information Processing Systems 26 (NeurIPS)*, 2013, 3111–3119. DOI: 10.48550/arXiv.1310.4546.

[4] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, **5** (2017), 135–146. DOI: 10.1162/tacl_a_00051.

[5] J. Wieting, M. Bansal, K. Gimpel, K. Livescu, "From Paraphrase Database to Compositional Paraphrase Model and Back", *Transactions of the Association for Computational Linguistics*, **3** (2015), 345–358. DOI: 10.1162/tacl_a_00143.

[6] N. Reimers, I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, 3982–3992. DOI: 10.18653/v1/D19-1410.

[7] A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder, L. Weng, "Text and Code Embeddings by Contrastive Pre-Training", *arXiv preprint arXiv:2201.10005*, 2022. DOI: 10.48550/arXiv.2201.10005.

[8] T. Gao, X. Yao, D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.

[9] J. J. Bartholdi, C. A. Tovey, M. A. Trick, "Voting schemes for which it can be difficult to tell who won the election", *Social Choice and Welfare*, **6**:2 (1989), 157–165. DOI: 10.1007/BF00303169.

[10] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, S. Belongie, "Generalized Non-metric Multidimensional Scaling", *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007, 11–18.

[11] A. M. Kolosov, A. I. Maysuradze, “Improving the quality of vector representations of words by using several sources of representations”, *Intelligent Systems: Theory and Applications*, **30**:1 (2026), 87–100.

[12] J. G. Kemeny, “Mathematics without numbers”, *Daedalus*, **88**:4 (1959), 577–591.

[13] A. H. Copeland, “A “reasonable” social welfare function”, *Seminar on Applications of Mathematics to Social Sciences, University of Michigan*, 1951.

[14] J.-C. Borda, “Mémoire sur les élections au scrutin”, *Histoire de l’Académie Royale des Sciences*, 1781, 657–665.

[15] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Sasson, G. Wolfman, E. Ruppin, “Placing Search in Context: The Concept Revisited”, *Proceedings of the 10th International Conference on World Wide Web*, 2001, 406–414. DOI: 10.1145/371920.372094.

[16] E. Bruni, N.-K. Tran, M. Baroni, “Multimodal Distributional Semantics”, *Journal of Artificial Intelligence Research*, **40** (2014), 1–47. DOI: 10.1613/jair.4135.

[17] F. Hill, R. Reichart, A. Korhonen, “SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation”, *Computational Linguistics*, **41**:4 (2015), 665–695. DOI: 10.1162/COLI_a_00237.

Received on May 15, 2026