УДК 004.85

Беляев Никита Александрович
студент, Национальный исследовательский ядерный университет «МИФИ», Россия, Москва
Юдина Диана Сергеевна
студент, Национальный исследовательский ядерный университет «МИФИ», Россия, Москва
Мачулин Владимир Владимирович
студент, Национальный исследовательский ядерный университет «МИФИ», Россия, Москва
Лапин Владислав Сергеевич
студент, Национальный исследовательский ядерный университет «МИФИ», Россия, Москва

КОНТЕКСТНО-ЗАВИСИМОЕ И МУЛЬТИМОДАЛЬНОЕ РАСПОЗНАВАНИЕ ИНТЕНСИЙ РЕЧЕВЫХ АКТОВ НА ОСНОВЕ ПРЕДОБУЧЕННЫХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

РЕЗЮМЕ

Цель. Разработать и экспериментально оценить систему автоматического распознавания речевых интенсий, учитывающую содержание анализируемой реплики, её положение в диалоге и акустические характеристики речи. Исследование направлено на проверку влияния диалогового контекста и совместного использования текстовой и речевой модальностей на качество классификации коммуникативных функций высказываний.

Материалы и методы. Сформирована система из 40 социально-реляционных интенсий и подготовлен двуязычный корпус, содержащий 8240 англоязычных и 2800 русскоязычных реплик из художественных диалогов. Разметка выполнялась комбинированным способом: первичное назначение класса большими языковыми моделями дополнялось автоматической проверкой согласованности и ручной валидацией неоднозначных примеров.

Текстовый классификатор построен посредством тонкой настройки предобученной модели DistilBERT. Вход модели включал анализируемое высказывание, предшествующий и, при наличии, последующий контекст. Для оценки вклада звучащей речи дополнительно переразмечены 1116 реплик корпуса Russian Emotional Speech Dialogs. На общей выборке сопоставлены текстовая, речевая на основе Wav2Vec2 и мультимодальная модели. Качество оценивалось по доле правильных ответов, Macro F1 и Weighted F1. Проверка разметки и предсказаний также проводилась с участием 19 респондентов.

Результаты. Включение соседних реплик повысило среднюю долю правильных ответов текстового классификатора с 70,8 до 77,3 %, то есть на 6,5 процентного пункта. Согласие респондентов с метками корпуса составило 83,3 %, а с предсказаниями модели на новых высказываниях — 78 %. В мультимодальном эксперименте текстовая модель достигла 70,4 % по accuracy, 64,8 % по Macro F1 и 68,8 % по Weighted F1; речевая модель — соответственно 64,1, 56,1 и 60,9 %. Наилучший результат показало объединение текстовых и акустических представлений: 82,3 % по accuracy, 72,9 % по Macro F1 и 79,1 % по Weighted F1. Преимущество мультимодальной модели над текстовой было статистически значимым по всем трём метрикам.

Выводы. Диалоговый контекст является существенным источником информации при распознавании речевых интенсий. Акустические признаки недостаточны для самостоятельной классификации, однако дополняют лексико-семантическое представление и позволяют точнее различать близкие коммуникативные действия. Разработанный подход может применяться в чат-ботах, виртуальных ассистентах и системах анализа обращений.

Ключевые слова: речевая интенсия, речевой акт, обработка естественного языка, трансформер, DistilBERT, Wav2Vec2, мультимодальная классификация, контекст диалога, машинное обучение.

ABSTRACT

Objective. To develop and evaluate a speech act intension recognition system

that uses the target utterance, dialogue context, and acoustic speech characteristics, and to test whether combining textual and speech modalities improves classification.

Materials and methods. A taxonomy of 40 socially oriented intensions was constructed. A bilingual corpus of 8,240 English and 2,800 Russian utterances from literary dialogues was prepared. Annotation combined labeling by large language models, automated consistency checks, and manual validation of ambiguous instances. The text classifier was implemented by fine-tuning DistilBERT and received the target utterance with preceding and, when available, following context. To assess spoken delivery, 1,116 utterances from the Russian Emotional Speech Dialogs dataset were re-annotated. Text-only, Wav2Vec2-based audio-only, and multimodal models were compared using identical data splits. Performance was measured with accuracy, Macro F1, and Weighted F1. Labels and predictions were additionally evaluated by 19 respondents.

Results. Adding adjacent dialogue turns increased mean text-classifier accuracy from 70.8% to 77.3%, an absolute gain of 6.5 percentage points. Respondents agreed with corpus labels in 83.3% of cases and with predictions for unseen utterances in 78.0% of cases. The text-only model achieved 70.4% accuracy, 64.8% Macro F1, and 68.8% Weighted F1; the audio-only model achieved 64.1%, 56.1%, and 60.9%, respectively. The combined model produced the best results: 82.3% accuracy, 72.9% Macro F1, and 79.1% Weighted F1. Its advantage over the text-only model was statistically significant for all three metrics.

Conclusions. Dialogue context is essential for recognizing speech act intensions. Acoustic characteristics are insufficient as an independent signal but complement lexical-semantic representations and improve discrimination between similar communicative actions. The approach can be used in chatbots, virtual assistants, educational platforms, and user-request analysis systems.

Keywords: speech act intension, natural language processing, transformer, DistilBERT, Wav2Vec2, multimodal classification, dialogue context, machine learning.

1. Введение

Современные диалоговые системы должны определять не только буквальное содержание пользовательского сообщения, но и действие, которое человек совершает посредством высказывания. Реплика может использоваться для просьбы, согласия, возражения, поддержки, уточнения, предложения компромисса или завершения разговора. При этом одна и та же последовательность слов в разных ситуациях способна выполнять разные коммуникативные функции.

В настоящей работе для описания такой функции используется понятие речевой интенсии. Интенсия представляет социально-реляционный компонент значения речевого акта и характеризует отношение говорящего к собеседнику, направленность высказывания и совершаемое коммуникативное действие [1–3]. Интенсию необходимо отличать от внутреннего намерения человека и эмоциональной окраски сообщения. Внутренние цели говорящего не всегда доступны наблюдению, тогда как выраженное в реплике коммуникативное действие может быть отнесено к определённому классу.

Автоматическое распознавание интенсий осложняется контекстной зависимостью. Например, реплика «Хорошо» может выражать добровольное согласие, вынужденную уступку, нейтральное подтверждение или намерение завершить обсуждение. Текстовые признаки также не всегда отражают иронию, степень уверенности, настойчивость и другие особенности произнесения. Поэтому перспективным является совместное использование текстового содержания, диалогового контекста и акустических признаков.

Научная новизна исследования состоит в сочетании контекстного представления реплики, специализированной системы из 40 социально-реляционных интенсий и мультимодального анализа текстовых и акустических признаков.

2. Цель исследования

Цель исследования — разработать и экспериментально оценить систему распознавания речевых интенсий на основе предобученных текстовых и

речевых моделей с учётом диалогового контекста.

Для достижения цели решались следующие задачи: формирование нормализованного списка речевых интенсий; подготовка корпуса русскоязычных и англоязычных реплик; разработка контекстно-зависимого текстового классификатора; оценка качества разметки и предсказаний с участием респондентов; сравнение текстовой, речевой и мультимодальной моделей.

3. Материалы и методы

3.1. Формирование системы интенсий

Исходный список классов формировался на основе исследований интенциональной структуры диалога, экспертных формулировок и результатов генерации вариантов с помощью больших языковых моделей. К формулировкам предъявлялись следующие требования: интенсия должна описывать одно речевое действие, иметь социальную направленность, не содержать конкретного ситуативного контекста и не объединять несколько независимых действий.

Первоначально было получено около 1000 вариантов. После удаления частных, дублирующихся и некорректно сформулированных элементов сформирован список из 125 классов. Затем близкие по смыслу классы были объединены. Итоговая система включила 40 интенсий, среди которых: просьба о совете, признание ошибки, предложение помощи, выражение согласия, вежливое несогласие, поддержка собеседника, предложение компромисса, уточнение информации и завершение разговора.

Для каждой интенсии подготовлены русскоязычная и англоязычная формулировки, имеющие одинаковую семантическую структуру.

3.2. Формирование текстового корпуса

Источниками реплик стали художественные произведения на русском и английском языках. Выбор художественных диалогов обусловлен тем, что они содержат прямые и косвенные речевые акты, неоднозначные высказывания и

контекстно зависимые коммуникативные ситуации.

Разметка выполнялась в три этапа. На первом этапе реплика вместе с предшествующим и последующим контекстом передавалась большой языковой модели, которая выбирала один класс из заданного списка. На втором этапе метка проверялась несколькими моделями или независимыми запусками модели-судьи. Примеры с несогласованными решениями направлялись на ручную проверку. На третьем этапе исключались реплики с недостаточным контекстом, выраженной многозначностью, дублированием или отсутствием определённой основной интенсии.

Каждая запись корпуса включала анализируемую реплику, метку интенсии, предшествующий контекст и, при наличии, последующий контекст.

Таблица 1 — Состав двуязычного корпуса

Язык	Число реплик	Число классов
Английский	8240	40
Русский	2800	40
Всего	11 040	40

3.3. Текстовая модель

Для текстовой классификации использовалась модель DistilBERT — компактная версия двунаправленной трансформерной модели BERT (Bidirectional Encoder Representations from Transformers) [4–6]. Применение предобученной модели позволяет использовать языковые закономерности, сформированные на больших текстовых корпусах, и адаптировать их к прикладной задаче посредством тонкой настройки.

Входная последовательность формировалась в формате: [CTX] предшествующий контекст [UTT] анализируемая реплика [AFT] последующий контекст. Специальные маркеры позволяют модели различать контекст и целевое высказывание. После токенизации последовательность обрабатывалась трансформерным кодировщиком. Полученное векторное представление передавалось в линейный классификационный слой, формирующий логиты для 40 классов.

Данные разделялись на обучающую, валидационную и тестовую части с сохранением распределения классов. При обучении применялись взвешенная кросс-энтропия, сглаживание меток, оптимизатор AdamW, dropout и ранняя остановка. На начальном этапе параметры языковой части модели замораживались и обучалась классификационная голова. Затем слои DistilBERT размораживались для полной тонкой настройки.

Программная реализация классификаторов выполнена с использованием библиотеки Transformers [7] и фреймворка PyTorch [8].

3.4. Речевая и мультимодальная модели

Для оценки вклада акустической информации использовался корпус Russian Emotional Speech Dialogs (RESD) [9]. Исходный набор содержит русскоязычные эмоциональные диалоги актёров и текстовые расшифровки. Для решаемой задачи 1116 реплик были дополнительно размечены по интенсиям.

В эксперименте исследовались три режима: text — классификация только по расшифровке; audio — классификация только по звуковому сигналу; text_audio — совместное использование обеих модальностей.

Текстовые признаки формировались трансформерной моделью, а речевые — предобученной моделью Wav2Vec2 [10]. В мультимодальном варианте векторные представления текста и аудиосигнала объединялись и передавались в общий классификационный слой. Для всех трёх вариантов использовалось одинаковое разделение на обучающую, валидационную и тестовую выборки.

3.5. Метрики оценки

Качество оценивалось с помощью accuracy (доли правильно классифицированных примеров), precision (точности), recall (полноты), Macro F1 и Weighted F1. Macro F1 вычисляет среднее значение F1 по классам без учёта их частоты и отражает качество распознавания редких интенсий. Weighted F1 учитывает объём каждого класса и характеризует качество при фактическом распределении данных.

Дополнительно анализировались матрица ошибок, зависимость качества

от длины реплики и влияние контекста. Для независимой проверки проведён эксперимент с участием 19 респондентов. Участники оценивали соответствие меток примерам корпуса и правильность предсказаний модели на новых репликах.

4. Результаты

4.1. Качество текстового классификатора

На основном корпусе средние значения по исследованным классам составили: accuracy — 0,75; F1 — 0,73; recall — 0,72; precision — 0,76.

Качество существенно различалось между классами. Высокие показатели были получены для согласия с собеседником и признания собственной ошибки. Наиболее сложным оказался класс, связанный с использованием юмора: его F1 составил 0,22. Это объясняется зависимостью юмористической функции от фоновых знаний, контекста, интонации и скрытого подтекста.

Дополнительный эксперимент показал выраженное влияние контекста. Без соседних реплик среднее значение accuracy составляло 70,8 %, а при добавлении контекста увеличивалось до 77,3 %. Абсолютный прирост равен 6,5 процентного пункта.

Наименьшая accuracy наблюдалась для очень коротких высказываний — 63,3 %. Для реплик средней длины показатель достигал 83,3 %. При дальнейшем увеличении длины качество снижалось, что может быть связано с появлением нескольких смысловых компонентов или нескольких интенсий внутри одного высказывания.

4.2. Экспертная оценка

При проверке разметки корпуса средняя доля согласия участников с назначенными метками составила 83,3 %. При оценке предсказаний модели на новых высказываниях согласие составило 78 %.

Снижение результата на новых репликах объясняется большей неоднозначностью примеров и зависимостью их коммуникативной функции от отсутствующего контекста. Вместе с тем показатель 78 % свидетельствует, что

модель способна переносить выявленные закономерности на высказывания, не использовавшиеся при формировании основного корпуса.

4.3. Сравнение модальностей

Результаты моделей на переразмеченной части корпуса RESD представлены в таблице 2.

Таблица 2 — Сравнение текстовой, речевой и мультимодальной моделей

Модель	Accuracy, %	Macro F1, %	Weighted F1, %
Текстовая	70,4	64,8	68,8
Речевая	64,1	56,1	60,9
Мультимодальная	82,3	72,9	79,1

Речевая модель показала худший результат. Акустические характеристики позволяют определить интонацию, темп и эмоциональную подачу, однако не содержат полной информации о предметном и прагматическом содержании реплики.

Мультимодальная модель превысила текстовую по всем метрикам: accuracy — на 11,90 процентного пункта; Macro F1 — на 8,05 процентного пункта; Weighted F1 — на 10,30 процентного пункта.

При парном сравнении accuracy разность составила 11,90 п.п. при 95%-м доверительном интервале от 5,40 до 18,40 п.п. и p-значении 0,004. Для Macro F1 получен интервал от 1,15 до 14,95 п.п. при p-значении 0,038, для Weighted F1 — от 3,80 до 16,80 п.п. при p-значении 0,006.

Следовательно, акустические признаки не заменяют текстовую информацию, но содержат дополнительный сигнал, позволяющий точнее различать близкие по формулировке речевые действия.

5. Обсуждение результатов

Полученные результаты подтверждают, что распознавание интенсии нельзя свести к классификации изолированной последовательности слов. Существенный прирост при включении контекста показывает, что коммуникативная функция определяется положением реплики в диалоге. Это особенно характерно для коротких ответов, косвенных просьб, формального

согласия, иронии и завершения разговора.

Отдельное использование речевых признаков оказалось менее эффективным, чем анализ текста. Данный результат закономерен: по темпу или интонации можно определить эмоциональную подачу, но сложно установить конкретное речевое действие без знания произнесённых слов. В то же время объединение модальностей даёт устойчивый прирост. Текст определяет основную семантику, а аудиосигнал уточняет отношение говорящего, степень уверенности, настойчивость и эмоционально-прагматический оттенок.

Практическая значимость разработанного подхода заключается в возможности включения классификатора в чат-боты, виртуальных ассистентов, образовательные платформы и системы анализа обращений. Определение коммуникативной функции позволяет выбирать реакцию системы не только по теме сообщения, но и с учётом характера пользовательского обращения.

Например, фразы, описывающие одинаковую проблему, могут выражать просьбу о помощи, требование, жалобу или запрос разъяснения. Реакция диалоговой системы должна различаться для каждой из этих ситуаций.

6. Выводы

Разработан подход к распознаванию интенсий речевых актов, объединяющий подготовку специализированного корпуса, контекстное представление текста и мультимодальную классификацию.

1. Сформирована система из 40 интенсий и двуязычный корпус из 11 040 размеченных реплик.
2. Использование диалогового контекста повысило accuracy текстового классификатора с 70,8 до 77,3 %, то есть на 6,5 процентного пункта. Согласие респондентов с разметкой корпуса составило 83,3 %, а с предсказаниями модели на новых высказываниях — 78 %.
3. Наилучший результат продемонстрировала мультимодальная модель: accuracy — 82,3 %, Macro F1 — 72,9 %, Weighted F1 — 79,1 %. Акустические признаки не заменяют текстовую информацию, но дают дополнительный сигнал для различения близких коммуникативных

функций.

4. Дальнейшие исследования следует направить на расширение корпуса спонтанной речью, уточнение границ между близкими классами, исключение утечек между выборками, разработку многометочной постановки и проверку моделей в независимых предметных областях.

Список литературы

1. Samsonovich A. V., Khabarov D. L., Belyaev N. A. Semantic Space Embedding of Speech Act Intensions // Optical Memory and Neural Networks (Information Optics). 2025. Vol. 34, Suppl. 1. P. S1–S15.
2. Гребенщикова Т. А., Зачесова И. А. Психология повседневного дискурса: интенциональный аспект. М.: Изд-во «Институт психологии РАН», 2014. 208 с.
3. Афиногенова В. А. Интенциональная организация повседневного диалога // Вестник Тверского государственного университета. Серия: Педагогика и психология. 2012. № 9. Вып. 2. С. 257–265.
4. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.
5. Devlin J., Chang M.-W., Lee K. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. Vol. 1. P. 4171–4186.
6. Sanh V., Debut L., Chaumond J. et al. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. 2019. arXiv:1910.01108.
7. Wolf T., Debut L., Sanh V. et al. Transformers: State-of-the-Art Natural Language Processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. P. 38–45.
8. Paszke A., Gross S., Massa F. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems. 2019. Vol. 32. P. 8024–8035.

9. Amentes A., Davidchuk N., Lubenets I. Russian Emotional Speech Dialogs with Annotated Text [Электронный ресурс]: набор данных. 2022. URL: https://huggingface.co/datasets/Aniemore/resd_annotated (дата обращения: 24.07.2026).
10. Baevski A., Zhou H., Mohamed A. et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 12449–12460.