

Новосибирский Государственный Университет

Теория аргументации и исследование методов формирования корпусов данных и обнаружения фейков на основе аргументации

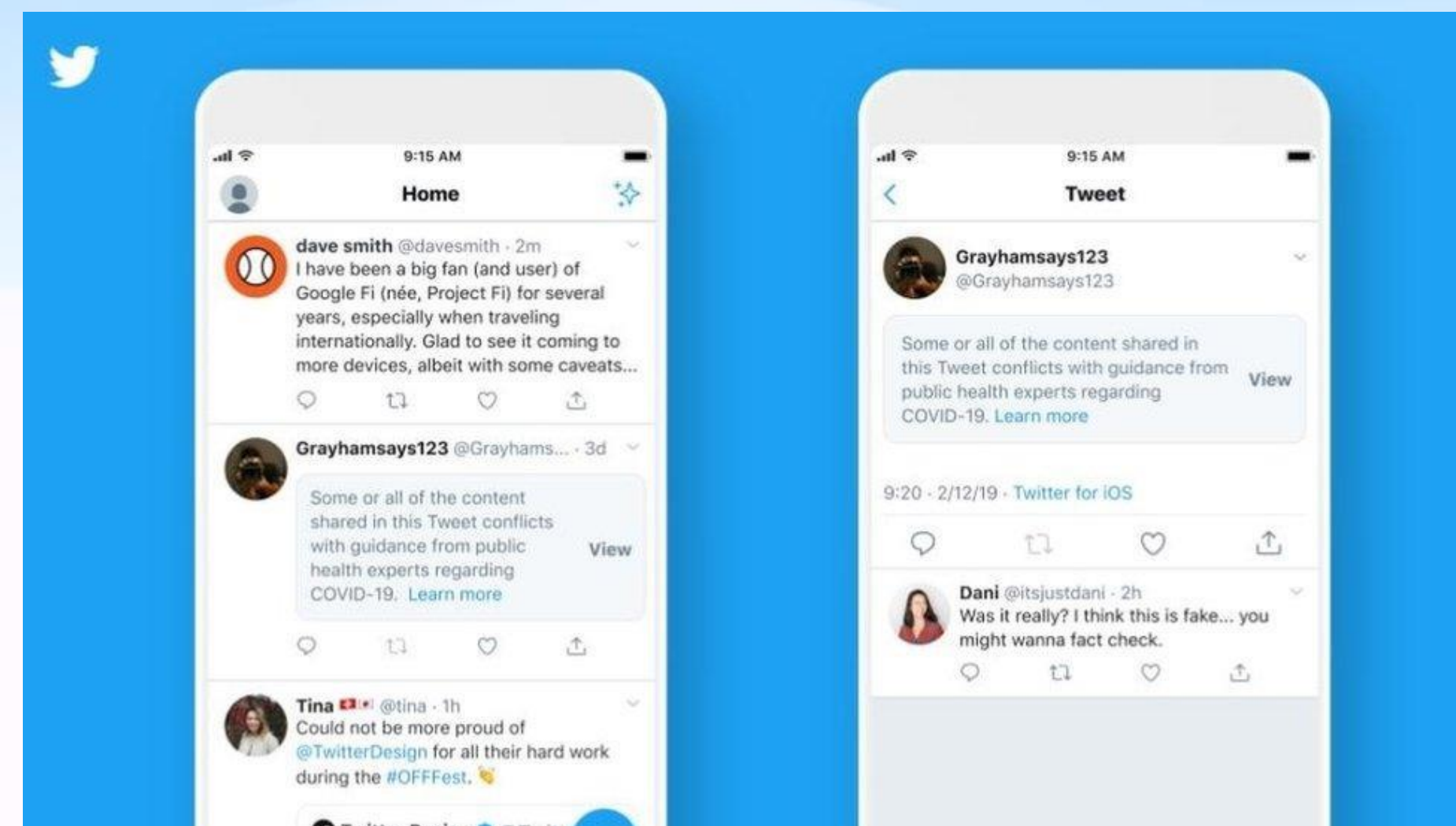
Кравченко Тимофей Максимович 24225.1

Руководитель: Сидорова Е.А., доцент каф. СИ ФИТ

Новосибирск 2025

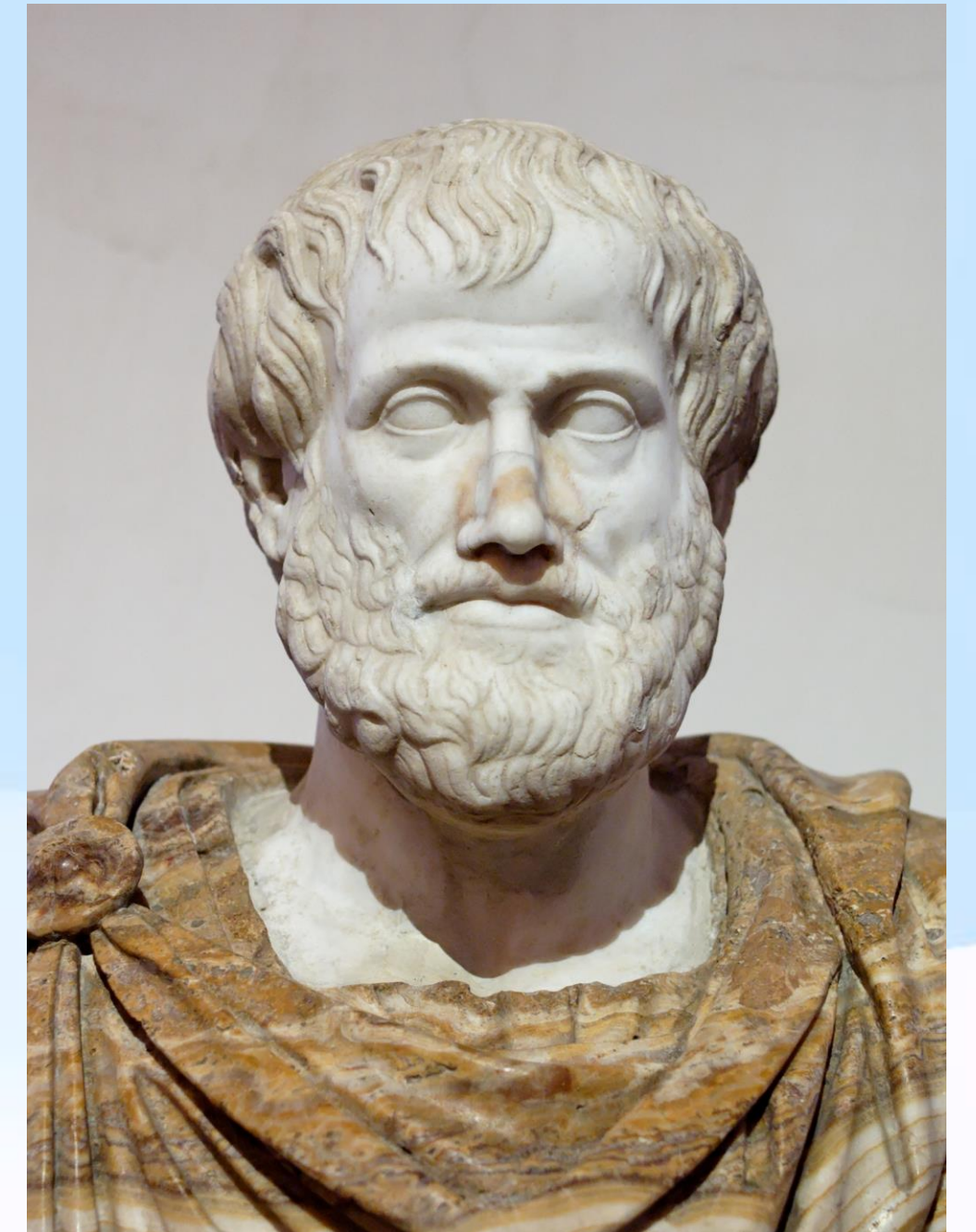
Актуальность

- В области автоматического анализа аргументативного дискурса естественного языка идентификация логических ошибок (фальсификаций) играет важную роль, поскольку может быть определяющим признаком для оценки качества аргументации. Кроме того, автоматическое выявление логических ошибок может быть полезно для развития систем детекции дезинформации и инструментов развития критического мышления. Обнаружение логических ошибок можно использовать не только для борьбы с дезинформацией, но и в юридической сфере или в обучении, чтобы помочь людям думать более критически. Огромную роль автоматический анализ играет и в политических кампаниях, климатических риториках и в медицине.
- По данным исследования, В 2023 году количество копий фейков увеличилось на 25% по сравнению с 2022 годом. На конец 2023 года ожидается рост количества копий фейков до 12,5 млн при сохранении тренда.



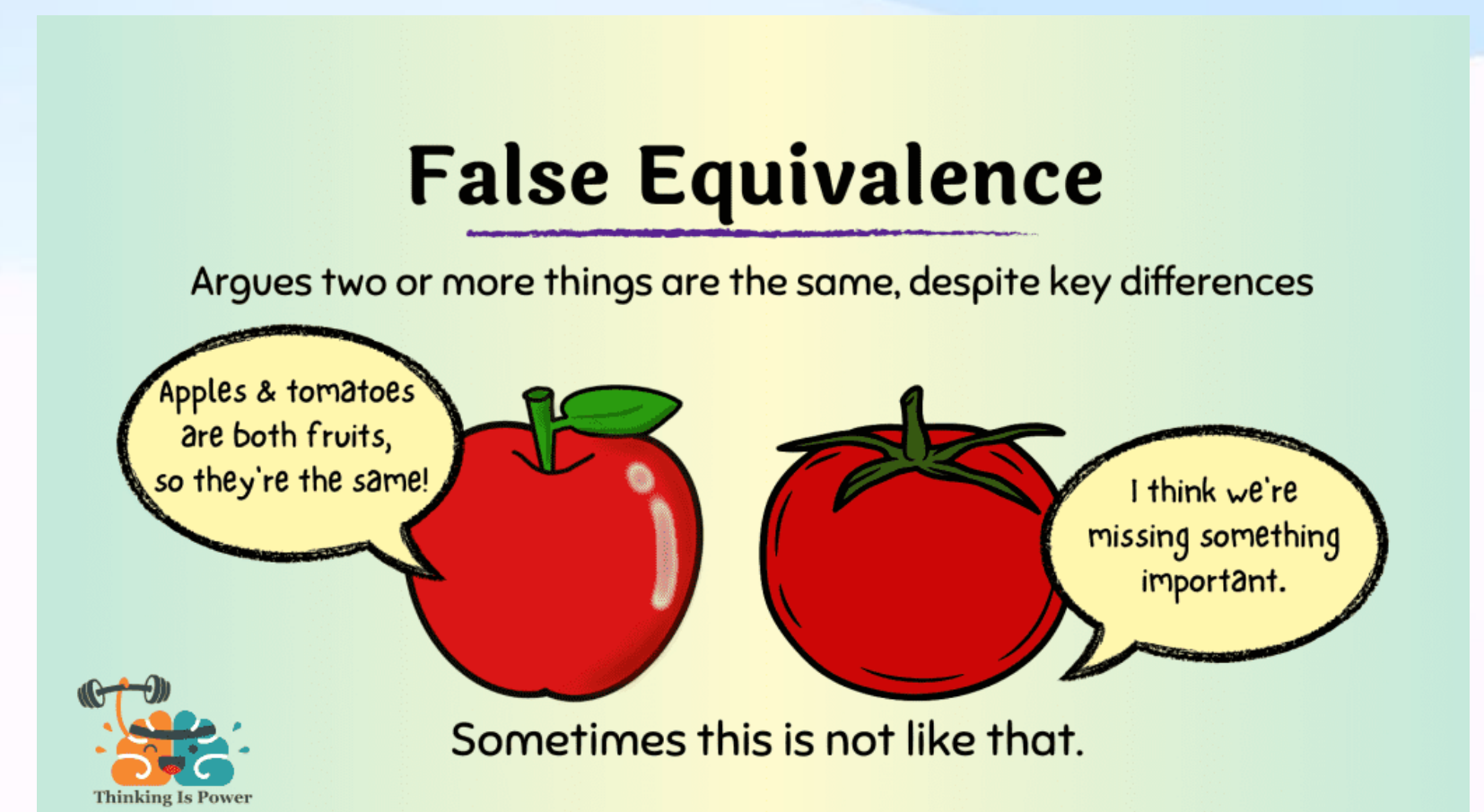
Введение

- Заблуждения изучаются со времен Древней Греции, ещё Аристотель рассматривал их как стратегию аргументации, используемую для обмана оппонента в споре и нечестного достижения победы.
- Что такое заблуждение (англ. fallacy)? Согласно определению Дуглас Нил Уолтона (1995), заблуждение - это ключевые виды ошибок или уловки в аргументации, которые часто вводят в заблуждение или запутывают участников в обычных повседневных спорах и обсуждениях.



Как распознать заблуждение?

- Обнаружение фрагмента ошибочного рассуждения не является тривиальной задачей и требует знаний в широком спектре областей, что делает эту задачу сложной. Во-первых, важно уметь анализировать логическое рассуждение, лежащее в основе аргументов на естественном языке. Для этого необходимо выделить абстрактные и формальные компоненты из неформального аргумента на естественном языке.
- Во-вторых, твердые знания в области обсуждения имеют первостепенное значение. Аргумент может быть логически корректным, но при этом ошибочным, как в случае неформальных ошибок (Walton, 1987). Поэтому только с помощью полного анализа возможно определить, является ли аргумент на естественном языке ошибочным, а также выявить причины, по которым он является таковым.
- Последовательный способ проведения такого анализа — это использование проверенных моделей аргументации, которые охватывают понятие ошибки. В литературе были предложены и изучены различные модели; одним из примеров является прагма-диалектическая теория аргументации (Van Eemeren и Grootendorst, 2016), в которой авторы определяют десять правил для направления аргументативных дискуссий. Выполнение этих правил позволяет создать плодотворное обсуждение, но аргумент, нарушающий любое из этих правил, считается ошибкой.



- **Правило свободы:**

Участники не должны препятствовать друг другу в выдвижении или оспаривании точки зрения.

- **Правило бремени доказательства:**

Сторона, выдвигающая точку зрения, обязана ее защитить, если об этом попросят.

- **Правило точки зрения:**

Нападки на точку зрения должны относиться к той точке зрения, которая действительно была высказана другой стороной.

- **Правило релевантности:**

Точка зрения может защищаться только с помощью аргументов, которые имеют отношение к этой точке зрения.

- **Правило невыраженных посылок:**

Участники не должны приписывать друг другу невыраженные посылки или отрицать ответственность за свои собственные невыраженные посылки.

- **Правило исходных позиций:**

Участники дискуссии не могут отказываться от соглашений, касающихся исходных позиций дискуссии.

- **Правило валидности:**

Рассуждения должны быть логически корректными или такими, которые могут быть сделаны корректными путем уточнения одной или нескольких невыраженных посылок.

- **Правило аргументативной схемы:**

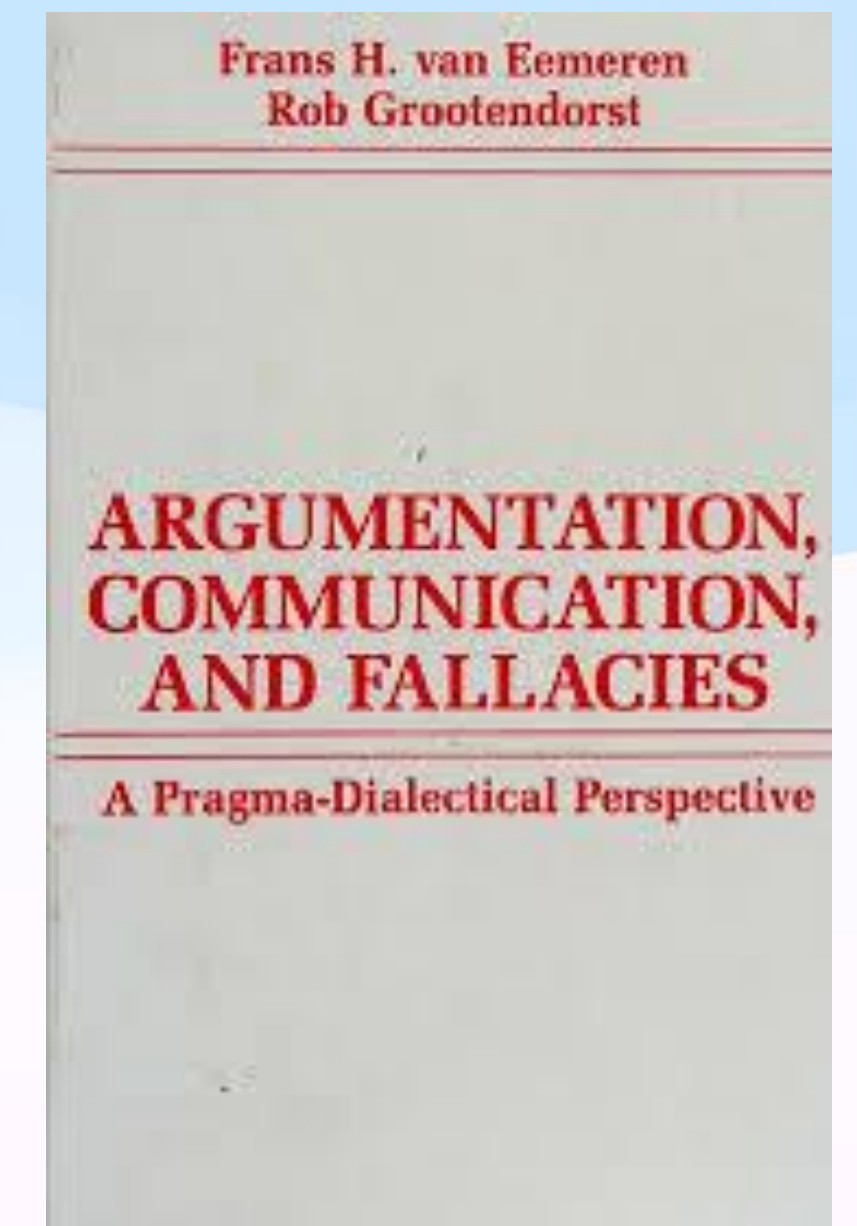
Точка зрения должна защищаться с использованием подходящих схем аргументации, которые применяются корректно.

- **Правило завершения:**

Неудачная защита точки зрения должна привести к отказу стороны от этой точки зрения, а успешная защита должна привести к снятию сомнений другой стороной.

- **Правило ясности:**

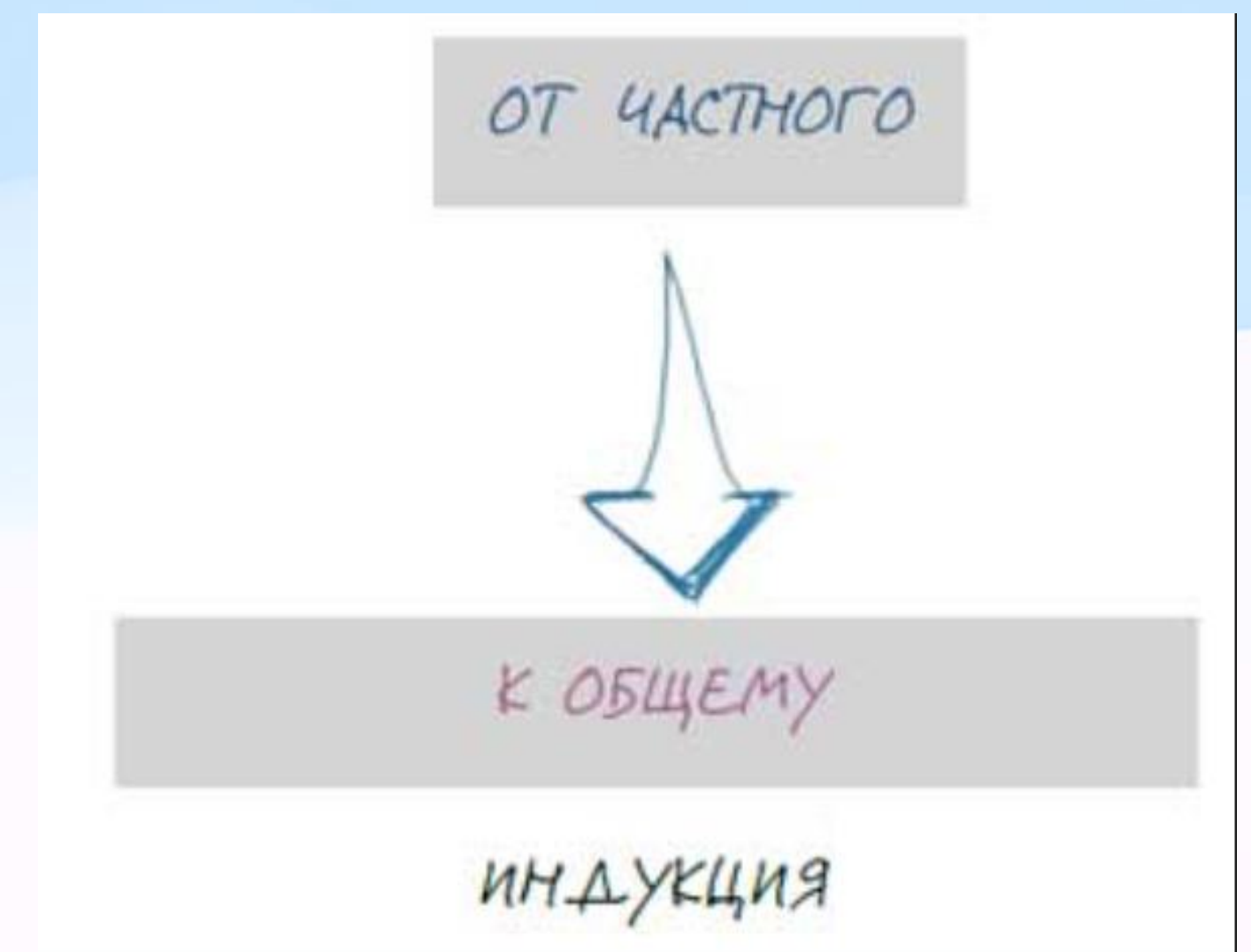
Участники не должны использовать формулировки, которые неясны, двусмысленны или запутанны, и должны интерпретировать аргументы другой стороны максимально четко и точно.



Аргументация

Индукция

- Индуктивные рассуждения извлекают вероятную (но не определенную) предпосылку из конкретных и ограниченных наблюдений. Есть данные, а затем из них делаются выводы — это называется индуктивной логикой.
- Пример: я вижу светлячков на своем заднем дворе каждое лето, гипотеза: этим летом я, вероятно, увижу светлячков на своем заднем дворе



Аргументация

Дедукция

- При использовании аргументации в тех или иных областях, зачастую обращаются к дедуктивному подходу.
- В дедуктивном рассуждении есть первая предпосылка, затем вторая предпосылка и, наконец, вывод (основанный на рассуждениях и доказательствах).
- Простой пример: “у всех пауков восемь ног. Тарантул-это паук. Следовательно, у тарантулов восемь ног”. Распространенной формой дедуктивных рассуждений является силлогизм, который основывается на следующем правиле: если у нас есть дизъюнкция (логическое "или") двух утверждений, и одно из этих утверждений ложно, то другое должно быть истинно.
- Проблема: должна строиться на истинных предпосылках, что не всегда представляется возможным, поэтому существуют иные методы рассуждений, о которых говорят авторы статьи.



Аргументация

Предположительная аргументация и опровержимая аргументация

- Предположительная аргументация - в логике и риторике относится к аргументам, которые основываются на предположениях или допущениях, а не на абсолютно достоверных фактах. Этот тип аргументации часто используется, когда прямые доказательства отсутствуют или когда необходимо сделать вывод на основе имеющихся данных
- Опровержимая аргументация — это тип аргументации, который основывается на предположениях, которые могут быть опровергнуты или пересмотрены при появлении новых фактов или доказательств. В отличие от дедуктивной логики, где выводы обязательно следуют из посылок, в опровержимой аргументации выводы считаются правдоподобными, но не абсолютными.

История построения аргументативных схем

- Perelman and Olbrechts-Tyteca, in *The New Rhetoric* (1969) - выделили и определили множество различных видов аргументов, используемых для убеждения собеседника на предварительной основе
- Arthur Hastings' Ph.D. thesis - предложил более систематическую классификацию, перечислив множество из этих схем вместе с полезными примерами, что послужило предпосылкой к Toulmin warrant (модель Тулмина)
- Eemeren and Grootendorst (1984; 1992). Kienpointner (1992) - разработали всеобъемлющее описание схем аргументации, которое включает дедуктивные и индуктивные, а также предположительные схемы.
- (Walton, 1996) - разработал список предположительных схем аргументации

Аргументационные схемы Уолтона

Argument from Position to Know (Version I)

- “Major Premise: Source a is in a position to know about things in a
- certain subject domain S containing proposition A.
- Minor Premise: a asserts that A (in Domain S) is true (false).
- Conclusion: A is true (false).”

Когда участник диалога выдвигает аргумент и он соответствует указанным выше требованиям, то он несет определенный вес как предположение. Однако он может быть опровергнут посредством критических вопросов. Соответствующие аргументу из позиции знания три критических вопроса приведены в работе Уолтона (1996, с. 62).

Пример критических вопросов для данной схемы аргументации

- CQ1: Is a in a position to know whether A is true (false)?
- CQ2: Is a an honest (trustworthy, reliable) source?
- CQ3: Did a assert that A is true (false)?

Когда участник диалога выдвигает аргумент из позиции знания, оппонент может задать любой из этих трех критических вопросов. Как только вопрос задан, предположительный вес, который аргумент имел до этого, отменяется. Но если участник предоставляет приемлемый ответ на вопрос, вес восстанавливается.

Аргументационные схемы Уолтона

Appeal to Expert Opinion (Version I)

Major Premise: Source E is an expert in subject domain S containing proposition A.

Minor Premise: E asserts that proposition A (in domain S) is true (false).

Conclusion: A may plausibly be taken to be true (false).

Обращение к мнению эксперта представляет собой опровержимую форму аргументации (defeasible argumentation scheme), которая не должна восприниматься как не подлежащая сомнению. Существует естественная тенденция уважать эксперта, и поэтому нам трудно ставить под сомнение слова эксперта. Тем не менее, обращение к мнению эксперта лучше всего рассматривать как предмет критического вопроса. В работе Уолтона (1997, с. 223) предлагается шесть основных критических вопросов.

1. Expertise Question (Вопрос об экспертности): How credible is E as an expert source?
2. Field Question(Вопрос о сфере): Is E an expert in the field that A is in?
3. Opinion Question(Вопрос об утверждении): What did E assert that implies A?
4. Trustworthiness Question(Вопрос о надежности): Is E personally reliable as a source?
5. Consistency Question(Вопрос о согласованности): Is A consistent with what other experts assert?
6. Backup Evidence Question(Вопрос о подтверждающих доказательствах): Is A's assertion based on evidence?

Примеры заблуждений по Уолтану

- Ad hominem (атака на личность): Когда аргумент направлен против человека, а не против его аргумента.
- Appeal to emotion (обращение к эмоциям): Когда аргумент пытается манипулировать эмоциями, чтобы повлиять на решение, а не использует логику или факты.
- Red herring (красная сельдь): Когда аргумент намеренно отвлекает от основной темы, вводя новую, не связанную с обсуждением.
- Hasty generalisation (поспешное обобщение): Когда делается вывод на основе недостаточного количества данных или нерепрезентативной выборки.
- Irrelevant authority (нерелевантный авторитет): Когда аргумент опирается на мнение эксперта, который не является компетентным в данной области.

Требования к аргументации

Для автоматического определения аргументации необходимо выполнение 4 свойств

- (i) Уметь максимально захватывать неявные рассуждения автора,
- (ii) Быть удобными для машинной обработки,
- (iii) Быть полезными для прикладных задач
- (iv) Сохранять простоту аннотации человеком

Однако схема аннотации Уолтона требует от аннотаторов ручного создания предложений на естественном языке для неявных заполнителей слотов (т.е. свойства (ii) и (iv) не учитываются).

Шаблоны аргументов

- Аннотируемые шаблоны (аргументные шаблоны или АТ), которые позволяют представить рассуждения автора без необходимости ручного создания предложений на естественном языке, когда заполнители слотов неявны. Они захватывают базовое рассуждение между двумя аргументативными сегментами в аргументативном отношении (AR) с использованием двух типов причинных меток (например, PROMOTE и SUPPRESS).

Argument: *I took an **NLP class**, an advanced course in Stanford. I suggest not taking **further advanced courses** because they will hurt your **GPA**.*

(a) Fallacy Classification (Jin+ 2022; Sourati+ 2023; etc.) **Faulty Generalization**

(b) Argumentation Schemes (Walton 2008) - - -

Scheme: Argument from Consequence

Support → **Claim:** **A** should not be brought about.
A = *further advanced courses*

Premise: If **A** is brought about, bad consequences will occur.

Operationalized

(b') Argument Templates (Reisert+ 2018) ← - - -

Template: Argument from Consequence-AT-S3

Support → **Claim:** BAD(**A**)
A = *further advanced courses*

Premise P: **A** SUPPRESS **C**, GOOD(**C**)
C = *GPA*

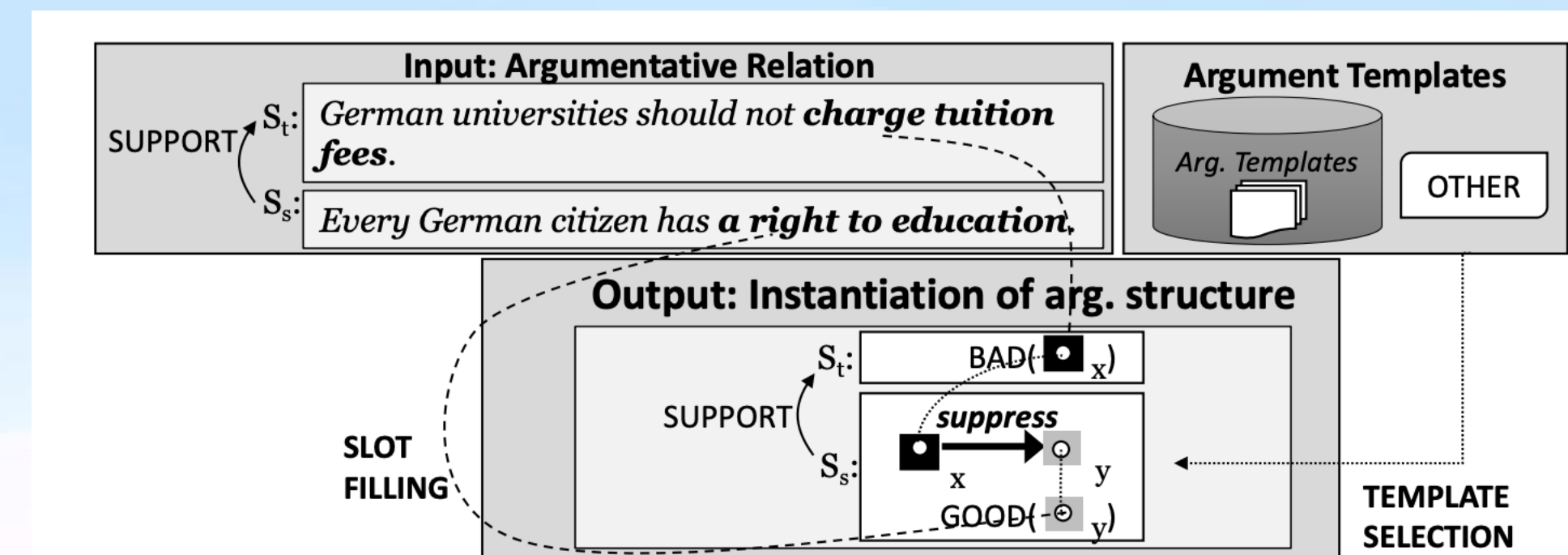
Шаблоны аргументов

Наша задача — выявить рассуждение между ними, выполнив два шага:

- (i) Выбрать шаблон из predetermined набора шаблонов (аргументные шаблоны (AT)), где каждый шаблон кодирует причинную метку
- (ii) Инстанцировать шаблон через заполнение слотов, где слот связан с релевантной, произвольной фразой в исходном тексте.

Первый шаг — идентификация AT: "S1, целевой сегмент отношения (т.е. St), утверждает, что x не должно быть совершено (т.е. плохо), потому что S2, исходный сегмент отношения (т.е. Ss), утверждает, что x плохо, потому что, когда x происходит, y, хороший объект/событие, будет подавлен".

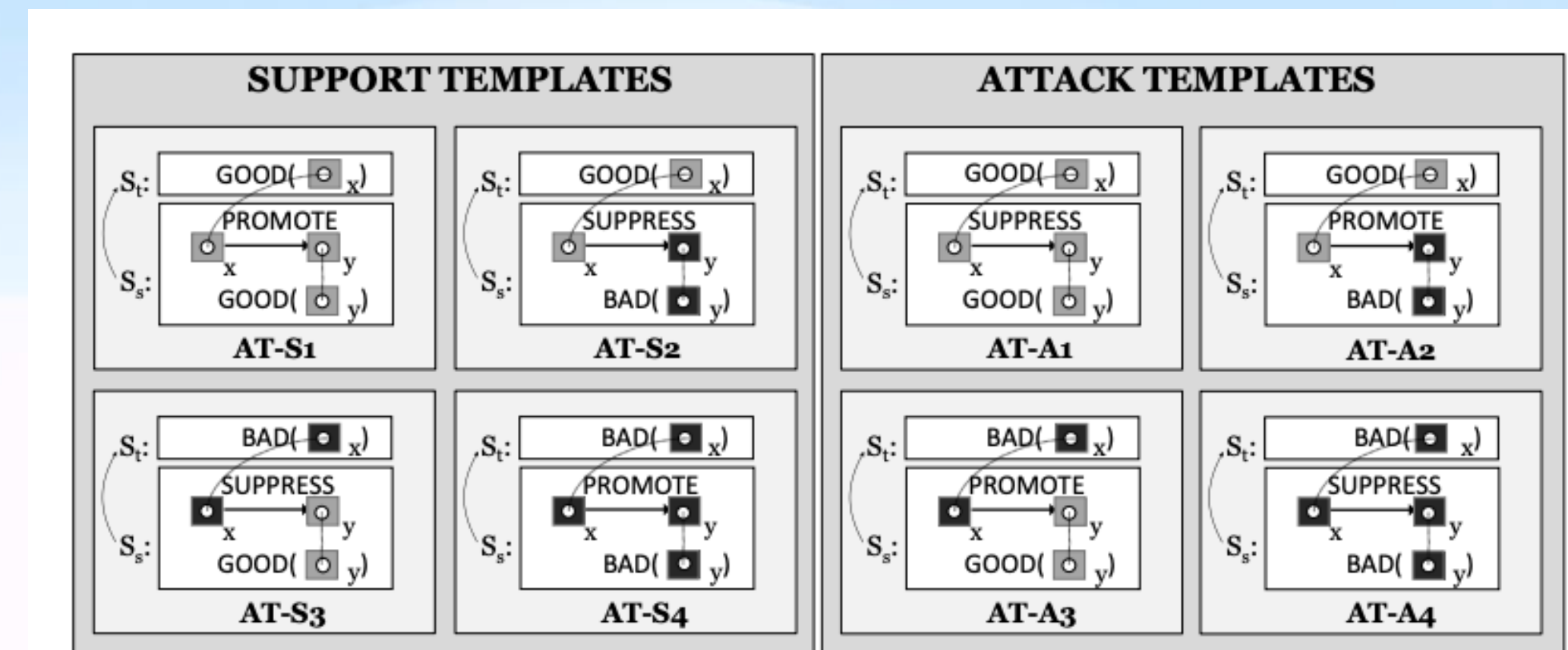
Второй шаг — инстанцирование шаблона путем заполнения слотов x и y фразами из текста: x = "взимать плату за обучение", y = "право на образование". Кодирова причинные метки, аннотаторам больше не требуется вручную конструировать неявные заполнители слотов



Шаблоны аргументов

Набор аргументов

- Набор аргументов (ATs) с использованием корпуса arg-microtexts (Peldszus и Stede, 2015a), который представляет собой корпус ручной компоновки аргументов. Корпус содержит 112 аргументативных текстов, каждый из которых состоит примерно из пяти сегментов, включающих вопрос по теме политики, основную претензию и несколько предпосылок. Каждый аргумент в тексте представляет собой политическое рассуждение, где каждая тема поддерживает или опровергает необходимость выполнения определенного действия. Кроме того, каждый аргументативный сегмент был аннотирован своей позицией (противник или сторонник) по отношению к вопросу темы. 357 аргументативных связей (ARs) между сегментами были вручную аннотированы как отношения SUPPORT (сегмент поддерживает приемлемость другого аргументативного сегмента), ATTACK (сегмент атакует приемлемость другого аргументативного сегмента) или UNDERCUT (сегмент атакует другую AR), где каждое отношение составляет 62,7% (224/357), 23,5% (84/357) и 13,8% (49/357) соответственно.



Шаблоны аргументов

Улучшенный вариант

Placeholder-ы (A, A', и C):

A: Основное множество действий или объектов (например, "продвинутые курсы").

A': Подмножество A, специфическое действие или объект (например, "курс по обработке естественного языка (NLP)").

C: Некоторая характеристика или результат, связанный с A (например, "высокий средний балл (GPA)").

Ошибочная генерализация: Ошибка возникает, когда A' (подмножество A) принимается за характеристику всего A. Например:

Если на основании одного конкретного курса (NLP) делают вывод о всех "продвинутых курсах".

Пример с гаражом:

A= "гараж".

C= "завышенные цены".

Логика: Гаражи в целом подозреваются в завышении цен (негативная характеристика C связана с A).

Однако рассматривается конкретный случай:

A'= "механик".

C'= "он завысил цену клиентке".

Вывод: Если механик завысил цену, это используется как доказательство того, что "все гаражи завышают цены".

Итоговая суть: Текст демонстрирует, как ошибки в аргументации появляются, когда свойства подмножеств (A' и C') необоснованно распространяются на все множество (A и C). Это имеет важное значение для анализа и понимания логических ошибок, а также предвзятых обобщений.

Argument: I took an **NLP class**, an advanced course in Stanford. I suggest not taking **further advanced courses** because they will hurt your **GPA**.

(a) Fallacy Classification (Jin+ 2022; Sourati+ 2023; etc.) **Faulty Generalization**

(b) Argumentation Schemes (Walton 2008) **Operationalized**

Scheme: Argument from Consequence

Support → **Claim:** **A** should not be brought about.
A = further advanced courses

Premise: If **A** is brought about, bad consequences will occur.

(b') Argument Templates (Reisert+ 2018)

Template: Argument from Consequence-AT-S3

Support → **Claim:** BAD(**A**)
A = further advanced courses

Premise P: **A** SUPPRESS **C**, GOOD(**C**)
C = GPA

(c) Fallacy Logic Structure (Our work)

Template: Faulty Generalization #2

Fallacious Support → **Premise P':** **A'** SUPPRESS **C**, GOOD(**C**)
A' = NLP class **C** = GPA
A' ⊆ **A**, but **A'** = **A** implicitly assumed

Faulty Generalization

Argument: My friend recently claimed that a **mechanic** at a certain **garage** **overcharged her**, and the evidence on her receipt seems to support this accusation. Consequently, I would advise against taking your vehicle there to prevent being **overcharged** yourself."

T1	T2
T3	T4
T5	

[A]= garage
[C]= overcharged
[A']= mechanic
[C']= overcharged her

Premise (P): [garage] PROMOTE [overcharged]
Premise (P'):

- [mechanic] PROMOTE [overcharged her]
- [mechanic] is a subset of [garage]
- [overcharged her] is a subset of [overcharged]

Conclusion: [garage] SHOULD NOT BE BROUGHT ABOUT

Виды шаблонов аргументации

Фокус исследования:

Это исследование концентрируется на анализе Аргумента от последствий, который часто используется, но подвержен ошибкам.

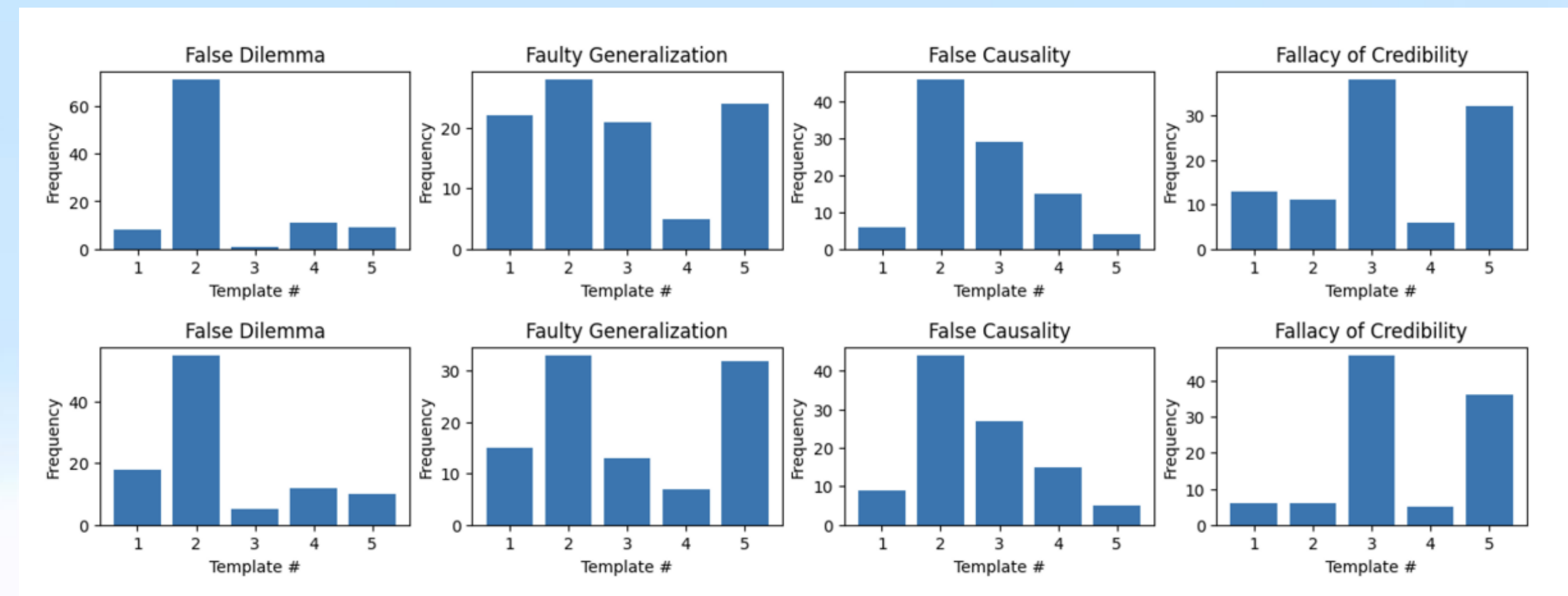
Для этого используются шаблоны аргументов (по примеру Reisert et al., 2018), чтобы выявить структуру логических ошибок (например, ложных выводов).

Такой подход помогает лучше анализировать распространённые ошибки в аргументах и предоставляет широкие возможности для аннотирования.

Fallacy of Credibility					
#1	A should not be brought about A — SUPPRESS —> GOOD(C) ↑ SUPPORT ↑ PROMOTE X	#2	A should be brought about A — SUPPRESS —> BAD(C) ↑ SUPPORT ↑ PROMOTE X	#3	A should be brought about A — PROMOTE —> GOOD(C) ↑ SUPPORT ↑ PROMOTE X
#4	A should not be brought about A — PROMOTE —> BAD(C) ↑ SUPPORT ↑ PROMOTE X	#5	No template can be instantiated		
False Causality					
#1	A should not be brought about A — SUPPRESS —> GOOD(C) ↑ SUPPORT ↑ RELATED TO A — SUPPORT —> GOOD(C)	#2	A should not be brought about A — PROMOTE —> BAD(C) ↑ SUPPORT ↑ RELATED TO A — SUPPORT —> BAD(C)	#3	A should be brought about A — PROMOTE —> GOOD(C) ↑ SUPPORT ↑ RELATED TO A — SUPPORT —> GOOD(C)
#4	A should be brought about A — SUPPRESS —> BAD(C) ↑ SUPPORT ↑ RELATED TO A — SUPPORT —> BAD(C)	#5	No template can be instantiated		
False Dilemma					
#1	¬A — SUPPRESS —> GOOD(C) ↓ SUPPORT A should be brought about ↑ SUPPORT A — PROMOTE —> GOOD(C)	#2	¬A — PROMOTE —> BAD(C) ↓ SUPPORT A should be brought about ↑ SUPPORT A — SUPPRESS —> BAD(C)	#3	¬A — PROMOTE —> GOOD(C) ↓ SUPPORT A should be brought about ↑ SUPPORT A — SUPPRESS —> GOOD(C)
#4	¬A — SUPPRESS —> BAD(C) ↓ SUPPORT A should be brought about ↑ SUPPORT A — PROMOTE —> BAD(C)	#5	No template can be instantiated		
Faulty Generalization					
#1	A should not be brought about A — SUPPRESS —> GOOD(C) ↑ SUPPORT ↑ SUPPORT A A' — SUPPRESS —> GOOD(C C')	#2	A should not be brought about A — PROMOTE —> BAD(C) ↑ SUPPORT ↑ SUPPORT A A' — PROMOTE —> BAD(C C')	#3	A should be brought about A — PROMOTE —> GOOD(C) ↑ SUPPORT ↑ SUPPORT A A' — PROMOTE —> GOOD(C C')
#4	A should be brought about A — SUPPRESS —> BAD(C) ↑ SUPPORT ↑ SUPPORT A A' — SUPPRESS —> BAD(C C')	#5	No template can be instantiated		

Аннотирование данных с помощью шаблонов аргументации

- GWET AC1 – метрика, используемая для измерения согласованности между аннотаторами, которая учитывает вероятность случайного совпадения.
- Krippendorff's α – универсальная метрика согласованности, применимая к различным типам данных (номинальным, интервальным и т.д.).
- Наибольшая согласованность:
 - False Causality (Ложная причинность): GWET AC1 = 0.71, Krippendorff's α = 0.65.
 - Это означает, что аннотаторы наиболее точно и согласованно классифицируют этот тип ошибки.
- Наименьшая согласованность:
 - Faulty Generalization (Ошибочная генерализация): GWET AC1 = 0.40, Krippendorff's α = 0.36.
 - Этот тип ошибки оказался самым сложным для согласованной классификации.
- Средние значения:
 - GWET AC1: 0.57.
 - Krippendorff's α : 0.54.
- В среднем согласованность между аннотаторами умеренная, но не высокая



Fallacy Type	GWET AC1	Krippendorff's α
False Dilemma	0.63	0.44
Faulty Generalization	0.40	0.36
False Causality	0.71	0.65
Fallacy of Credibility	0.58	0.49
Average	0.57	0.54

Синтетические данные

Основные моменты:

Подходы к генерации данных:

Zero-shot: gpt-3.5 генерирует примеры логической ошибки без предоставления примеров в подсказке (только описание типа ошибки и форма вывода — например, предложение, твит или вопрос-ответ).

Few-shot (1, 2, 5-shot): в подсказку добавляются 1–5 примеров логической ошибки, после чего модель генерирует такое же количество новых примеров.

Few-shot-context: Помимо примеров ошибки, подсказка включает дополнительный контекст (предыдущие и последующие предложения), и модель генерирует примеры вместе с контекстом.

Контроль качества данных:

Для каждого типа ошибки генерируется одинаковое количество примеров, чтобы изучить, как качество синтетических данных зависит от разных типов подсказок.

Исключаются частые и малоразнообразные ошибки, например:

Loaded Language (эмоционально нагруженные слова).

Name Calling (оскорбления или навешивание ярлыков), которые встречаются только в одном наборе данных (PROPAGANDA).

Prompt

Your task is to perform the following actions:

1. Read the sentence that has the fallacy of **Irrelevant Authority** and its provided context. The sentence and context below are both delimited by <>.
2. Generate a similar example of one sentence and its context of the **Irrelevant Authority** fallacy. The context should be a bigger chunk of text that includes the sentence, similar to the provided example sentence and context below.

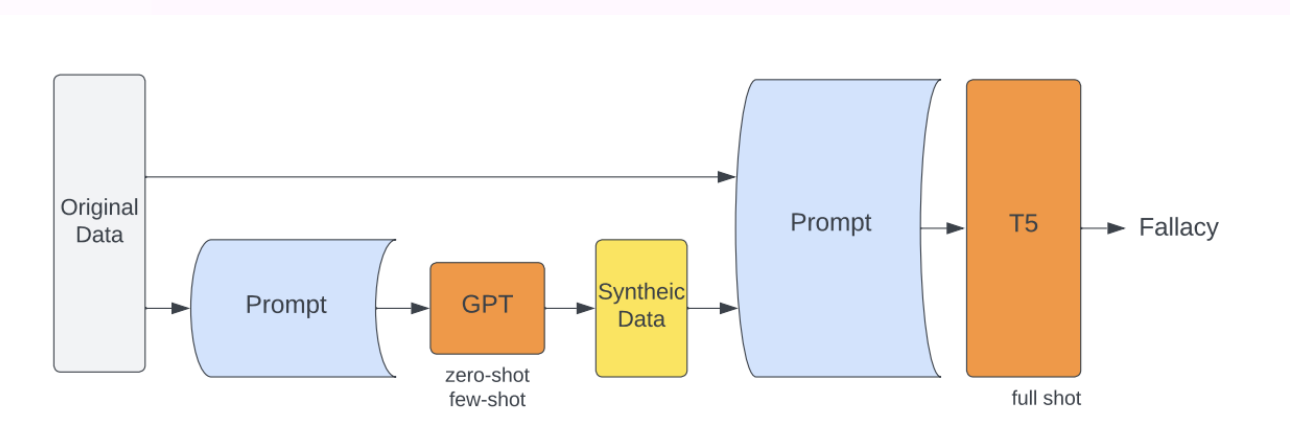
Your generated output should be in JSON format with the following keys: `generated_sentence`, `generated_context`.

Sentence: <Carlson cited Dr. Robert Epstein who has said, in Carlson's words, "Google alone could determine the outcome of almost any election just by altering its search selections and we would never know it.">

Context: <Carlson cited Dr. Robert Epstein who has said, in Carlson's words, "Google alone could determine the outcome of almost any election just by altering its search selections and we would never know it." Dr. Robert Epstein is a social scientist and an expert on Google. >

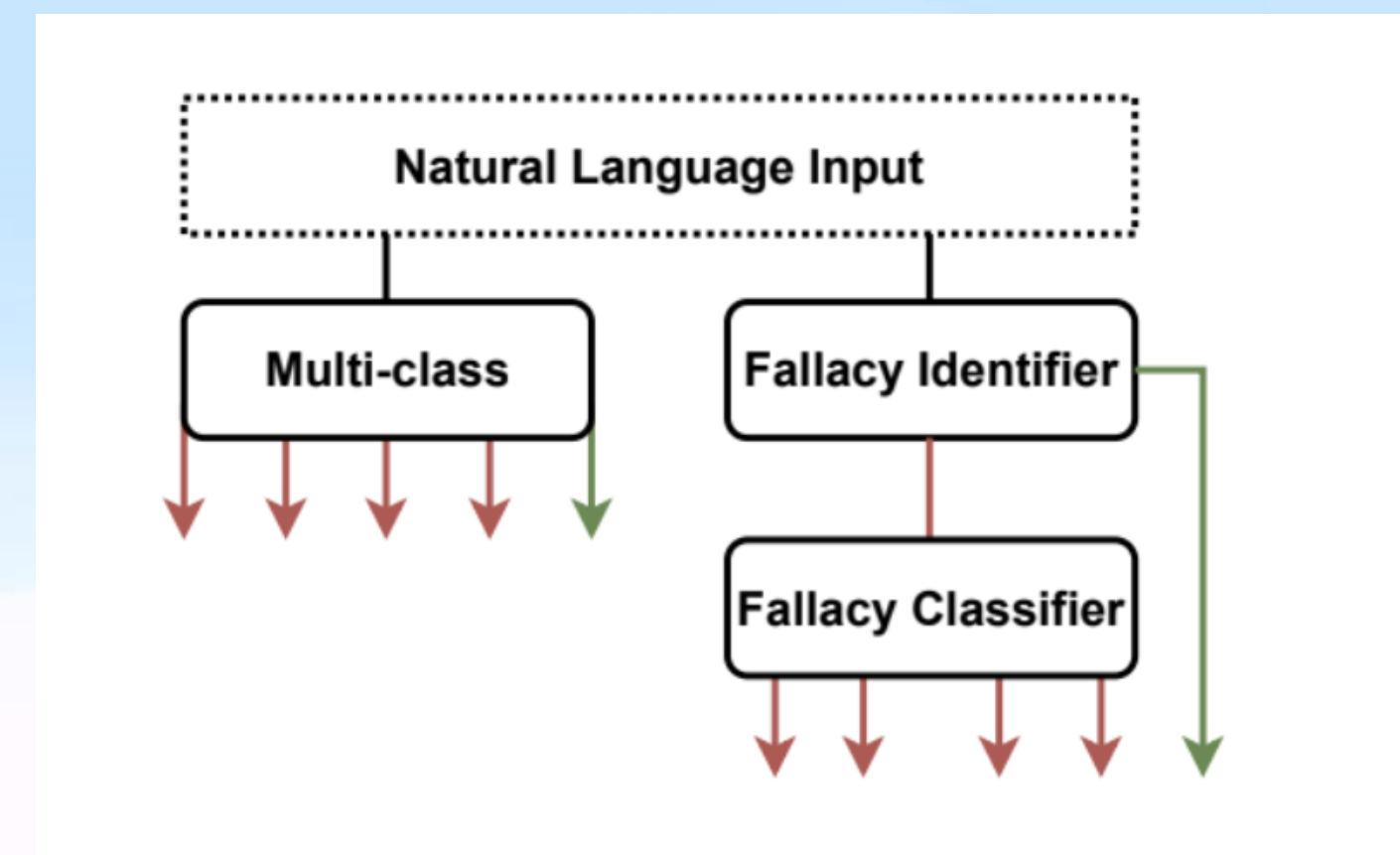
Generated Output

```
{  
  'generated_sentence': 'As a dog groomer, I can tell you that the earth is flat.',  
  'generated_context': 'While most people accept that the earth is round, one dog groomer is convinced otherwise. As a dog groomer, I can tell you that the earth is flat. Dogs are great navigators, and they always sense when the ground is flat or sloping.'  
}
```



Архитектуры систем для обнаружения фальсификации

- Классифицировать текст сначала на наличие фейков, а при наличии таковых, классифицировать их на подвиды фальсификации
- Производить многоклассовую классификацию



Model	Precision	Recall	Macro-F1
RB	21.6	24.6	18.6
eSVM	68.3	55.8	60.3
RoBERTa	68.2	65.3	66.5
GPT-3.5-TURBO	59.0	46.2	45.5
GPT-4	53.5	55.0	51.7
eSVM+[<i>ctx</i>]	67.3	50.0	54.4
RoBERTa+[<i>ctx</i>]	62.0	58.4	59.9
GPT-3.5-TURBO+[<i>ctx</i>]	50.2	32.1	35.8
GPT-4+[<i>ctx</i>]	54.4	51.2	50.8

Table 3: Precision, Recall and Macro-F1 results of the 5-class fallacy detection task. [*ctx*] represents the contextual information added to the input of each model.

Model	Precision	Recall	Macro-F1
RB	47.1	47.0	46.4
eSVM	77.8	77.5	77.7
RoBERTa	79.8	79.6	79.6
GPT-3.5-TURBO	41.7	46.2	40.6
GPT-4	53.2	53.2	51.1
eSVM+[<i>ctx</i>]	76.8	74.0	74.8
RoBERTa+[<i>ctx</i>]	78.0	78.8	78.3
GPT-3.5-TURBO+[<i>ctx</i>]	47.1	48.8	43.5
GPT-4+[<i>ctx</i>]	56.6	56.7	54.1

Table 4: Precision, Recall and Macro-F1 results of the 2-class fallacy detection task. [*ctx*] represents the contextual information added to the input of each model.

Model	Precision	Recall	Macro-F1
RB	22.9	22.1	22.4
eSVM	69.6	65.5	67.1
RoBERTa	75.4	78.0	76.2
GPT-3.5-TURBO	51.7	46.4	44.6
GPT-4	60.4	60.0	58.3
eSVM+[<i>ctx</i>]	79.7	72.1	74.8
RoBERTa+[<i>ctx</i>]	72.3	72.6	72.3
GPT-3.5-TURBO+[<i>ctx</i>]	45.9	38.1	35.1
GPT-4+[<i>ctx</i>]	58.7	57.0	55.7

Table 5: Precision, Recall and Macro-F1 results of the 4-class fallacy classification task. [*ctx*] represents the contextual information added to the input of each model.

Заключение

- Были изучены основные методы аргументации, а также их разновидности: аргументационные схемы, шаблоны аргументации; способы формирования синтетических данных и архитектуры систем для определения фальсификаций в текстах

ИСТОЧНИКИ

- Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition (<https://arxiv.org/pdf/2311.09552>)
- Flee the Flaw: Annotating the Underlying Logic of Fallacious Arguments Through Templates and Slot-filling (<https://arxiv.org/pdf/2406.12402v1#page=9&zoom=100,401,94>)
- Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models
- Argumentation Schemes and Defeasible Inferences
- Detecting Fallacies in Climate Misinformation: A Technocognitive Approach to Identifying Misleading Argumentation (<https://arxiv.org/pdf/2405.08254#page=13&zoom=100,133,508>)
- ArgNetBank Studio — платформа для моделирования и анализа АРГУМЕНТАЦИИ в дискурсе. (<https://uniserv.iis.nsk.su/arg/>)

Новосибирский Государственный Университет

Теория аргументации и исследование методов формирования корпусов данных и обнаружения фейков на основе аргументации

Кравченко Тимофей Максимович 24225.1

Руководитель: Сидорова Е.А., доцент каф. СИ ФИТ

Новосибирск 2025