СОФИЙСКИ УНИВЕРСИТЕТ
„СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО
МАТЕМАТИКА И ИНФОРМАТИКА

SOFIA UNIVERSITY
St. KLIMENT OHRIDSKI
FACULTY OF
MATHEMATICS AND INFORMATICS

Интелигентен агент за диалог на естествен език с отчитане на контекста

от

Момчил Емилов Хардалов

Автореферат

за присъждане на образователна и научна степен "Доктор"

в

професионално направление: 4.6 "Информатика и компютърни науки";
докторска програма: "Софтуерни Технологии" – Откриване на знания

Ръководител: проф. д-р Иван Койчев
Консултант: проф. д-р Преслав Наков

Октомври 2022

Дисертационният труд съдържа 179 страници, от които 8 страници са приложения. Включва 18 фигури и 39 таблици в 6 глави, общи заключения, приноси и 2 приложения. Библиографията обхваща 350 заглавия, всички от които на английски език. Списъкът на публикациите на автора, пряко отразяващи резултатите от дисертационния труд, съдържа 6 заглавия. Понастоящем има 77 известни цитата на тези публикации (намерени с Google Scholar).

Съдържание

Глава 1. Въведение

1.1 Мотивация и актуалност на темата

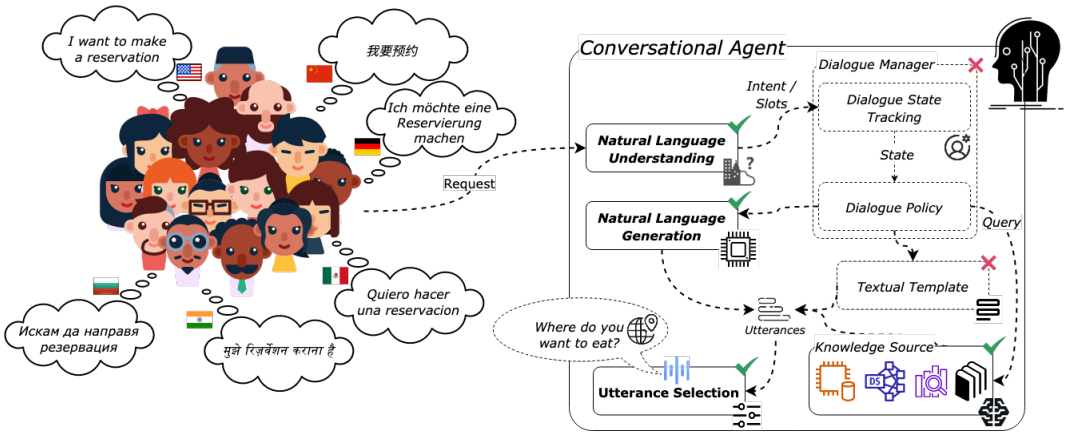
Интернет преобрази много области от нашето ежедневие. Той направи достъпни много нови услуги и продукти за хората по целия свят. В следствие на това се промени начина, по който компаниите и предприятията комуникират и взаимодействат със своите клиенти. Обръща се все повече внимание на качеството и надеждността при работа с клиенти, като от значение вече е не само коректността на информацията, но и времето, нужно за решение на конкретен проблем. От друга страна, тези услуги трябва да бъде достъпни за клиентите, както през предпочитаните от тях канали за комуникация, така и на предпочитания от тях език. Разговорът с хора, експерти, е по-вероятно да бъде по-пълноценен от клиентска гледна точка, но набирането и обучението на нови служители става все по-трудно. Това е скъп и времеемък процес, който не може да следва темпото на непрекъснато нарастващия брой нови потребители. Този дисбаланс, показва спешната нужда, от една страна, за допълнителна автоматизация с разговори агенти, и от друга, за разработване на по-надеждни системи за отговаряне на въпроси. Всички тези системи трябва да бъдат допълнени от нови и подобрени инструменти, които да бъдат използвани от операторите, обслужващи клиенти.

Първо, нека формално да дефинираме *разговорен агент*. Тази дефиниция ще бъде използвана в целия дисертационнен труд: "*Разговорен агент, наричан още чатбот, е компютърна програма, която се опитва да генерира отговори, подобни на тези, които човек би използвал по време на диалог.*" (Ramesh et al., 2017). Второ, следвайки Gao et al. (2019) ще определим обхвата на проблемите, които разговорните агенти се очаква да решат. Те могат да бъдат обобщени чрез следните три изследователски въпроса:

- **отговаряне на въпроси**: *"агентът трябва да предоставя кратки, директни отговори на потребителски запитвания въз основа на богати познания, извлечени от различни източници на данни, включително текстови колекции като уеб документи и предварително компилирани бази от знания, които съдържат данни за продажби и маркетинг*";
- **завършване на задачи**: "*агентът трябва да изпълнява потребителски задачи, вариращи от резервация на ресторант до насрочване на срещи и планиране на бизнес пътувания*";
- **социален чат**: "*агентът трябва да разговаря безпроблемно и по подходящ начин с потребителите – подобно на човек както в теста на Тюринг – и да предоставя полезни препоръки.*".

За да се определи актуалността на проблема, ще се фокусираме върху метрики и резултати опасани в няколко скорошни проучвания. Първо, важно е да се подчертае, че разговорните агенти печелят повече доверие както от компаниите, така и от самите клиенти, и се превръщат в неразделна част от техните канали за обслужване на клиенти. *Докладът за състоянието на разговорния маркетинг*

на Drift за 2020 г.[1] отбелязва, че използването на чатботове като комуникационен канал на търговските марки се е увеличил с 92% спрямо предходната година. В доклада на Zendesk[2] се отбелязва, че 69% от клиентите са готови да взаимодействат с бот по прости въпроси, което е 23% увеличение спрямо предходната година. Според Invesp 33% от потребителите биха предпочели да се свържат с отдела за обслужване на клиенти на компанията чрез социална медия, а не по телефона.[3] Въпреки това 54% от клиентите споделят, че най-голямото им разочарование от чатботовете е броят на въпросите, на които трябва да отговорят, преди да бъдат прехвърлени към оператор.[2] Освен това клиентите са притеснени от способностите на разговорните агенти да "разбират", 60% от тях смятат, че хората са в състояние да определят нуждите им по-добре от чатботовете.[4] Освен това потребителите отбелязват "неспособността на чатботовете да решават сложни проблеми" като друг техен недостатък.[5]



Фигура 1.1: Концептуална диаграма, илюстрираща потока на информация в **ориентиран към задачи разговорен агент**. Компонентите, които са изследвани в тази дисертация, са маркирани с ✓, а тези, които *не* са – с ✗.

На Фигура 1.1 са илюстрирани основните компоненти в един разговорен агент. Първият компонент, който обработва потребителската заявка, е този за *разбиране на естествен език (PEE) (natural language understanding)* (Weld et al., 2022). Той отговаря за цялостното разбиране на входния текст, следователно от там е и името на модула. Основните му задачи са (*i*) да открие намерението (*intent detection*) и (*ii*) да извлече стойностите за съответните слотове от входните токени (*slot filling*) и да ги предаде на *диалоговия мениджър* (*Dialogue Manager*). На свой ред *диалоговият мениджър* агрегира целия контекст на диалога, наречен проследяване на състоянието на диалога (dialogue state tracking) (Williams et al., 2016), определя целта на потребителя и генерира следващото системно действие, т.е. *политиката за диалог* (*dialogue policy*). Тук трябва да се отбележи, че тази дисертация не включва изследвания свързани с *диалоговия мениджър*. Основният ѝ фокус е върху подобряването на разбирането на естествен език – качеството на отговорите и генерираните изказвания (обсъдени по-долу) не само по отношение на тяхната фактологията, но и по отношение на съгласуваност и уместност спрямо потребителският вход.

[1] https://www.drift.com/blog/state-of-conversational-marketing/
[2] ,https://cx-trends-report-2022.zendesk.com/growth-areas
[3] https://www.invespcro.com/blog/social-media-customer-support/
[4] https://userlike.com/en/blog/consumer-chatbot-perceptions
[5] https://startupbonsai.com/chatbot-statistics/

Следващата стъпка от работата на разговорния агент е да свърже диалоговия акт (*dialogue act*), генериран от диалоговата политика, към изказване на естествен език (Gatt and Krahmer, 2018; Dong et al., 2022). За да се постигне това, често се прилагат множество стратегии като използване на модели за генериране на естествен език (ГЕЕ), попълване на предварително дефинирани текстови шаблони или извличане на данни от външни източници на знания. Шаблоните са неразделна част от ориентирания към задачи диалог (Williams and Zweig, 2016; Wen et al., 2017). Те конструират последователни и добре написани изречения, но имат същите проблеми като системите, базирани на правила, а именно, че са статични и изискват предварителна (ръчна) подготовка. Освен това агентът става по-ограничен, а диалогът звучи по-неестествено и по-еднообразно. Поради тези причини, шаблоните също не са включени в този дисертационен труд. В допълнение, моделите за ГЕЕ отговарят на потребителските въпроси, използвайки външни източници на знания, като например извличат дълги документи с отговори или намират пасажи с доказателства.[6]

Последната част от обработката е моделът за *избор на следващо изказване*. В случай че е използван един източник за генериране на естествен език (или подобен компонент), този модел трябва просто да копира предложения текст като следващ ход на чатбота, т.е. да бъде прескочен. Обаче, в случай че са използвани повече източници, разговорният агент трябва да избере най-подходящото изречение от списъка с кандидати. Така компонентът за *избор на следващо изказване* е отговорен за крайното подреждане (*re-ranking*) и избора на най-подходящата опция за следващ ход на агента. Решението отново може да бъде основано на предварително определен сценарий и правила. Тази дисертация изследва по-сложни методи, базирани на дълбоки невронни мрежи.

1.2 Цели и задачи на дисертационния труд

Целите на тази теза могат да бъдат обобщени, както следва:

1. Разработване на ефективни, базирани на обработка на естествен език, подходи за изграждане на многокомпонентни, ориентирани към задачите, отчитащи контекста, разговорни агенти със специфичното приложение като чатботове за обслужване на клиенти.

2. Създаване на нови ресурси и корпуси, които могат да помогнат при разработването на интелигентен агент за диалог, от една страна, разширявайки ги до множество езици, а от друга страна, позволявайки им да генерират разширени (дълги) отговори (напр. статии, извлечени от бази със знания), за разлика от обикновените къси такива.

Във връзка с това са очертани следните *задачи* на изследването:

- Преглед на съществуващата литература, предишни разработки и подходи за реализация на разговорни агенти и техните компоненти.

- Проектиране, описание, разработка и оценка на базиран на разбиране на естествен език (РЕЕ) компонент, който съвместно идентифицира потребителското намерение и разпознава какво е релевантно за неговите слотове.

[6] Клиентите предпочитат базите от знания пред всички други канали за самообслужване. https://www.hubspot.com/knowledge-base

- Проектиране, описание, разработка и оценка на алгоритъм за изготвяне на изказвания чрез външни източници на знания.
- Проектиране, описание, разработка и оценка на цялостни (*end-to-end*) генеративни модели за чатботове с приложение в клиентската поддръжката.
- Проектиране, описание, разработка и оценка на система за многоезичен и междуезичен диалог.

1.3 Структура на дисертацията

Останалата част от тази теза е организирана, както следва:

- В Глава 2 са разгледани най-съвременните подходи, свързани с агентите за диалог и техните компоненти. Първо са разгледани предишни разработки на ориентирани към задачи разговорни агенти – включително модулни и крайни (диференцируеми) диалогови системи. Второ, обхваща подходи, за решаване на две от основните задачи за разбиране на естествен език в ориентираният към задачите диалог – класификация на намеренията, запълване на слотове и тяхното съвместно моделиране. След това са разглеждани методи за отговаряне на въпроси (QA) и машинно четене с разбиране (МЧР), с фокус върху данни за отговаряне на научни въпроси, многоезични модели и подходи за междуезиков трансфер на знания. Следва обобщение на предишната работа по извличане на дълги обяснения през призмата на задачата за *откриване на предишни проверени твърдения*. И накрая са обсъдени усъвършенствани разговорни агенти, като цялостно (*end-to-end*) генериране, и стратегии за комбиниране на отговори от различни източници, напр. извлечени от предишни разговори, генерирани с помощта на модел тип последователност към последователност или чрез попълване на предварително дефинирани шаблони.
- В глава 3 е описан нов метод за съвместно откриване на намерение и запълване на слотове. Основната идея е да се използва по-добре връзката между двете задачи. За тази цел представянията на двете задачи се сливат заедно, докато се обучава моделът, от една страна, чрез механизъм за обединяване на вниманието (*attention pooling*), а от друга, чрез моделиране на слотове чрез свързване на представянията на ниво токен от езиковия модел с предвидено разпределение на намеренията и накрая добавяне на ръчно изработени функции. Освен това се демонстрират най-високи (*state-of-the-art*) резултати на два стандартни корпуси за РЕЕ, а именно ATIS (Hemphill et al., 1990) и SNIPS (Coucke et al., 2018).
- Глава 4 представя нови методи за извличане на отговори от външни източници на знания. Първо е описан нов корпус за отговаряне на въпроси с множествен избор на български език и оценка на методи, базирани на извличане на информация (information retrieval), за получаване на доказателствени пасажи. Освен това са изследвани методи за трансфер на знания без допълнително обучение (*zero-shot*) от богат на ресурси език (т.е. английски) към такъв с малко ресурси. След това е представен нов метод за получаване на дълги, описателни отговори, т.е. обяснения в контекста на откриване на предишни проверени твърдения. Предложен е нов метод с дистанционно наблюдение (*distant supervision*) за събиране на големи корпуси от двойки

статия-претенции и учене от тях с техники за самоадаптиране на модела към обучение върху шумни данни.

- В глава 5 са изследвани усъвършенствани методи за диалог. Първо са проучени цялостни генеративни агенти, обучени върху разговори в социалните медии, между оператор в компания, отговарящ за обслужване на клиенти и клиент. След това е представен нов подход за избор на отговор от множество източници, използвайки базиран на невронни мрежи модел за класиране. Накрая е представен нов многоезичен корпус за задачата за отговаряне на въпроси и са изследвани възможностите на няколко от най-съвременните многоезични модели за трансфер на знания между езици.

- Глава 6 завършва дисертацията, обобщава приносите и обсъжда посоки за бъдещи изследвания.

Научни публикации, свързани с дисертацията

- Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, AACL-IJCNLP '22, Online

- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020b. EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 5427–5444, Online

- Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2020a. Enriched Pre-trained Transformers for Joint Slot Filling and Intent Detection. *arXiv preprint arXiv:2004.14848*

- Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019a. Beyond English-Only Reading Comprehension: Experiments in Zero-shot Multilingual Transfer for Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '19, pages 447–459, Varna, Bulgaria

- Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019b. Machine Reading Comprehension for Answer Re-Ranking in Customer Support Chatbots. *Information*, 10(3)

- Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2018. Towards Automated Customer Support. In *Proceedings of the 18th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '18, pages 48–59, Varna, Bulgaria

Глава 2. Обзор на литературата

В тази глава се разглежда широк набор от цялостни подходи към разговорните агенти, включително корпуси използвани за обучение, и са представени основните понятия, които да послужат за контекст в останалата част от дисертацията.

Първо се обобщава литературата за две от задачите за PEE част от разговорните агенти: (*i*) откриване на намерение, т.е. разбиране на текущата цел на потребителя, и (*ii*) запълване на слот, т.е. идентифициране на различни слотове в текущия диалог, които съответстват на различни параметри на заявката на потребителя.

След това главата се фокусира върху задачата за отговаряне на въпроси (QA), обхващайки подходи за обучение върху целия корпус, както и трансфер без допълнително обучение, приложени в едно- и многоезичен сценарий. Освен това е разгледан проблемът за извличане на отговори от външни източници на знания. Той е разгледан през призмата на задачата за *откриване на вече проверени твърдения* като са описани най-съвременните подходи и модели, включително такива базирани на обучение върху шумни данни и дистанционно наблюдение (*distant supervision*).

И накрая на тази глава, фокусът е върху напредъка в техниките за провеждане на разговор, т.е. генеративни модели за диалог и комбиниране на отговори. Тези отговори могат да бъдат получени от множество източници, като целта е за да се намери най-доброто следващо изказване в разговор.

Глава 3. Семантичен анализ на потребителски изказвания

Тази глава представя нов метод за разбиране на естествен език, който моделира съвместно задачите за откриване на намерения и запълване на слотове (Transformer-NLU). Таблица 3.1 показва потребителска заявка, получена от личен гласов асистент. Тук намерението е да се *пусне музика* от *2005* година на изпълнителя *Justin Broadrick*. Задачата за запълване на слотовете възниква естествено като задача за тагиране на последователности. Основната идея на предложения в тази глава метод е, да се използва слой с обединяване на вниманието (*attention pooling*) за класифициране на намеренията, който използва цялостно представяне на входното изречение, образувано от всички токени. Тези векторни представяния също кодират информация за слотовете. Освен това задачата за попълване на слотовете е подпомогната от характеристики базирани на предсказанията за главни букви в думите, предсказаното разпределение от слоя за разпознаване на намеренията и други специфични характеристики. Те позволяват на модела да

различава име на хора, градове, държави, щати и т.н. Методът постига по-високи резултати спрямо най-съвременните модели базирани на многослойни невронни мрежи.

Intent	PlayMusic						
Words	play	music	from	2005	by	justin	broadrick
	↓	↓	↓	↓	↓	↓	↓
Slots	O	O	O	B-year	O	B-artist	I-artist

Таблица 3.1: Пример от корпуса SNIPS със слотове, кодирани в BIO формат. Намерението на изказването е *PlayMusic*, а дадените слотове са *year* и *artist*.

3.1 Корпус

В експериментите са използвани два публично достъпни корпуса (вижте Таблица 3.2), Airline Travel Information System (ATIS) (Hemphill et al., 1990) и SNIPS (Coucke et al., 2018). ATIS съдържа транскрипции от аудио записи на заявки за информация за полети, докато SNIPS е събиран от автоматична система за откриване на намерения използвана от персонални гласови асистенти.

	ATIS	SNIPS
Размер на речника	722	11,241
Средна дължина на изреченията	11.28	9.05
#Намерения	21	7
#Слотове	120	72
#Обучаващи примери	4,478	13,084
#Валидационни примери	500	700
#Тестови примери	893	700

Таблица 3.2: Статистика за корпусите ATIS и SNIPS.

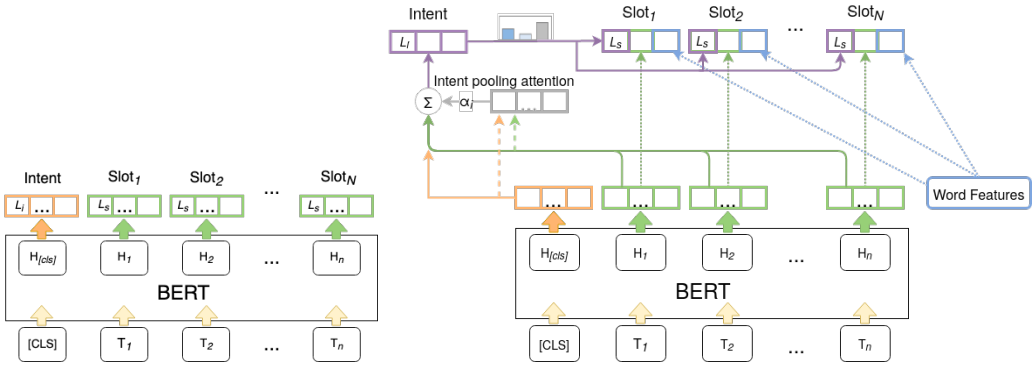
3.2 Метод

Предложеният съвместен метод за класификация на намеренията и запълване на слотове е изграден върху предварително обучен езиков модел, т.е. BERT (Devlin et al., 2019) или RoBERTa (Liu et al., 2019). В допълнение, базовият модел е подобрен по три начина: (*i*) за откриване на намерение се използва обединено представяне от последните скрити състояния за всички токени, (*ii*) допълнителни характеристики от предсказанията за главни букви в думите (*truecasing*) и наименуваните единици (*named entities*) за всеки токен, наречени характеристики на думата, и (*iii*) предсказаният вектор с вероятностното разпределение на разпознатите намерения, последните скрити представяния на BERT и характеристиките на думата се подават на слой за запълване на слот. Пълната архитектура на модела е показана на Фигура 3.1б.

За обучение на модела се използва функция на съвместна грешка $\mathcal{L}_{joint}$, която комбинира грешките на намерението и слотовете. И за двете задачи се прилага кръстосана ентропия (*cross entropy*) след софтмакс (*softmax*) активационен слой, освен в случай на тагиране с CRF. В тези експерименти функцията на грешката за слотовете $\mathcal{L}_{slot}$ е отрицателната логаритмична вероятност (*negative*

log likelihood) (NLL). Освен това е въведен нов хипер-параметър γ, за да се балансират грешките на двете задачи (вижте Формула 3.1). И накрая, загубата се разпространява (*propagate*) през всички немаскирани позиции в последователността, включително части от думи (*word pieces*) и специални токени ([CLS], <s> и т.н.). Трябва да се има предвид, че теглата на модела *не* се замразяват по време на допълнителното обучение (*fine tuning*).

$$\mathcal{L}_{joint} = \gamma * \mathcal{L}_{intent} + (1 - \gamma) * \mathcal{L}_{slot} \tag{3.1}$$



(а) BERT-Joint. **(б)** Transformer-NLU).

Фигура 3.1: Архитектури на модели за съвместно обучение на намерение и запълване на слотове: (a) класическо съвместно обучение с BERT/RoBERTa, и (b) предложена подобрена версия на модела.

3.3 Експерименти и Анализи

3.3.1 Оценка на модела

Таблица 3.3 представя резултати от количествената оценка на предложените модели по отношение на (*i*) точност на намерението, (*ii*) точност на изречението, и (*iii*) F1 на ниво слот. Първата част на таблицата се отнася за модели предложени в предишни изследвания, докато втората част представя извършените експерименти. Втората част е отделена с двойна хоризонтална линия. Резултатите от оценката потвърждават, че предложения модел се представя по-добре от текущите най-съвременни модели.

Въведена е нова мярка, а именно, процент на *относително намаляване на грешката* (RER), която се дефинира като пропорцията на абсолютната грешка, намалена от $model_a$ спрямо $model_b$.

$$RER = 1 - \frac{Error_{model_a}}{Error_{model_b}} \tag{3.2}$$

Таблица 3.4 показва намаляването на грешките от предложения модел в сравнение с текущите най-добри модели и базови такива основаващи се на BERT (вижте Раздел 3.4.2 в дисертацията). Тъй като няма единствен най-добър модел от най-съвременните такива, е взет максималният резултат от колоните за всички модели, въпреки че те не са получени чрез единно обучение. За корпуса ATIS виждаме намаление на грешката на точността на ниво изречение от 11,64% (1,49 в абсолютни точки) и 6,25% (0,25 в абсолютни точки) за F1 на ниво слот, но само

Model	ATIS			SNIPS		
	Intent (Acc)	Sent. (Acc)	Slot (F1)	Intent (Acc)	Sent. (Acc)	Slot (F1)
Joint Seq. (Hakkani-Tür et al., 2016)	92.60	80.70	94.30	96.90	73.20	87.30
Atten.-Based (Liu and Lane, 2016)	91.10	78.90	94.20	96.70	74.10	87.80
Sloted-Gated (Goo et al., 2018)	95.41	83.73	95.42	96.86	76.43	89.27
Capsule-NLU (Zhang et al., 2019)	95.00	83.40	95.20	97.30	80.90	91.80
Interrelated SF-First (E et al., 2019)	97.76	86.79	95.75	97.43	80.57	91.43
Interrelated ID-First (E et al., 2019)	97.09	86.90	95.80	97.29	80.43	92.23
Stack-Propagation (Qin et al., 2019)	96.9	86.5	95.9	98.0	86.9	94.2
AGIF (Qin et al., 2020)	97.1	87.2	96.0	98.1	87.3	94.8
BERT-Joint	97.42	87.57	95.74	98.71	91.57	96.27
RoBERTa-Joint	97.42	87.23	95.32	98.71	90.71	95.85
Transformer-NLU:BERT	**97.87**	**88.69**	**96.25**	**98.86**	91.86	**96.57**
Transformer-NLU:RoBERTa	97.76	87.91	95.65	**98.86**	**92.14**	96.35
Transformer-NLU:BERT w/o Slot Features	97.87	88.35	95.97	98.86	91.57	96.25
Transformer-NLU:BERT w/ CRF	97.42	88.26	96.14	98.57	92.00	96.54

Таблица 3.3: Резултати от откриване на намерения и запълване на слотове в корпуси SNIPS и ATIS. Най-високите резултати във всяка категория са написани с **удебелен шрифт**. Предложените модели са показани в *курсив*; моделите без курсив отгоре идват от литературата. Qin et al. (2019, 2020) докладват резултатите си с единична точност.

4,91% за точността на намерението (вижте Таблица 3.3). Такова малко подобрение може да се дължи на качеството на данните и на техния размер. За корпуса SNIPS виждаме значително подобрение на всички мерки и над 35% намаляване на грешките. В абсолютно изражение – 0,76 за намерение, 4,84 за изречение и 1,77 за слотове. Тези ефекти не могат да да бъдат отдадени само на по-добрия модел (обсъден в анализа по-долу), но също така и на имплицитната информация, която BERT научава по време на обширното си предварително обучение. Това е особено полезно в случая на SNIPS, където голяма част от слотовете в категории като *SearchCreativeWork, SearchScreeningEvent, AddToPlaylist, PlayMusic* са имена на филми, песни, изпълнители и т.н.

Metric	Relative Error Reduction	
	ATIS	
Intent (Acc)	4.91%	17.44%
Sent. (Acc)	11.64%	11.43%
Slot (F1)	6.25%	19.87%
	SNIPS	
Intent (Acc)	40.00%	11.63 %
Sent. (Acc)	35.91%	6.76%
Slot (F1)	37.64%	17.35%
Transformer-NLU	vs. най-съвременни	vs. BERT

Таблица 3.4: Сравнение по отношение на относително намаляване на грешка (Формула 3.2) на *Transformer-NLU:BERT* с два вида базови модели: *(i)* най-високите резултати на текущите най-съвременните модели (*SOTA*) за всяка метрика и *(ii)* стандартно допълнително обучен BERT-Joint (без промяна на архитектурата).

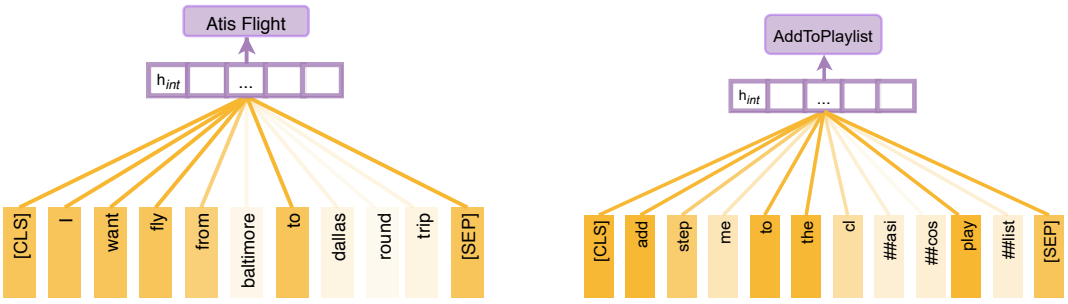
3.3.2 Анализ на Transformer-NLU

За да се изследва предложения модел и в частност да се определи количественият принос на неговите компоненти, те се добавят или премахват последователно. Резултатите са показани във втората част на Таблица 3.3. Първо са сравнени резултатите от *BERT-Joint* и обогатения модел *Transformer-NLU:BERT*. Забелязваме видимо намаляване на грешката при класифициране на намеренията съответно със 17,44% и 11,63% за корпусите ATIS и SNIPS. Освен това виждаме намаляване на грешките от 19,87% (ATIS) и 17,35% (SNIPS) за запълването на слот (F1) и 11,43% (ATIS) и 11,63% (SNIPS) за точност на изречението. Когато се използва RoBERTa като основа на предложения модел, въпреки че все още се забелязва положителния ефект от предложената архитектура, общите резултати са малко по-лоши. Това може да бъде отдадено на различния корпус, използван при предварителното обучение (CommonCrawl срещу Уикипедия). По-нататък анализът е фокусиран върху моделите, базирани на BERT, тъй като те се представят по-добре от тези базирани на RoBERTa.

В следващите експерименти се премахват допълнителните характеристики при запълването на слотовете – предсказаното намерение, регистъра на думите и именуваните единици. Резултатите са показани като *Transformer-NLU:BERT w/o Slot Features*. Както се очаква, точността на намерението остава непроменена и за двата корпуса, тъй като е запазен слоя за обединяване на вниманието (*attention pooling*), докато F1-резултатът за слотовете намалява. За SNIPS моделът постига същия резултат като *BERT-Joint*, докато при ATIS разликата е в рамките на 0,2 абсолютни точки.

Добавен е CRF слой след мрежата за запълване на слотове, заради показаните положителни ефекти в предишни методи (Xu and Sarikaya, 2013; Huang et al., 2015; Liu and Lane, 2016; E et al., 2019). Експериментът е означен като *Transformer-NLU:BERT w/ CRF*. Тук не се забелязва очакваното подобрение. Резултатите за запълване на слотовете са близки до най-високите регистрирани, в същото време се наблюдава драстичен спад в точността на откриване на намерение, т.е. -17,44% за ATIS и -20,28% за SNIPS.

Последно, на Фигура 3.2 са визуализирани научените тегла на вниманието. Те са извлечени от заявки от корпусите ATIS (Фигура 3.2а) и SNIPS (Фигура 3.2б).



(а) atis_flight (ATIS). **(б)** AddToPlaylist (SNIPS).

Фигура 3.2: Теглото на вниманието от слоя за обединяване на намерение за един пример от всеки корпус. Колкото по-плътна е линията, толкова по-голямо е теглото на вниманието.

3.4 Обобщение

В тази глава бяха изследвани двете от основните задачи в ориентираното към задачи, разговорно разбиране на естествен език, т.е. откриване на намерение и запълване на слотове. Те представляват важна част (компонент) от чатботовете за обслужване на клиенти, отговарящи на потребителски заявки. Тези заявки могат да бъдат получени през уебсайта на компанията или на различни корпоративни уеб платформи или социални медии. Компонент отговаря за извличането на двойки слот-стойност, които по-късно се използват от *диалоговия мениджър* за навигация в следващите действия на агента.

Беше предложен обогатен, предварително обучен езиков модел за съвместно моделиране на двете задачи, а именно, *Transformer-NLU*. Проектиран е слой за обединяване на вниманието с цел получаване на представяне на намерението извън извлеченото от специалния начален токен. Освен това беше подобрено запълването на слотовете със специфични за думата характеристики и предсказаното разпределение на намеренията. Чрез експерименти върху два стандартни корпуса беше показано, че Transformer-NLU превъзхожда другите алтернативи по всички стандартни мерки, използвани за оценка на PEE задачи. Беше показано, че използването на RoBERTa и добавянето на CRF слой върху мрежата за запълване на слотове не носят допълнително подобрение на метриките. И накрая, Transformer-NLU:BERT постигна точност за откриване на намерение от 97,87 (ATIS) и 98,86 (SNIPS), или като относително намаляване на грешка – почти 5% за ATIS и над 40% за SNIPS в сравнение с най-съвременните модели. По отношение на F1 за запълване на слотовете, предложените модели постигнаха резултати от 96,25 (+6,25%) за ATIS и 96,57 (+37,64%) за SNIPS.

Глава 4. Извличане на отговори от външни източници на знания

Тази глава предлага различни подходи за избиране на отговори от външни източници на знания. Тук фокусът е върху методи, които разчитат на допълнителна информация за да се отговори на потребителски въпрос (или заявка). Това може да е извличане на контекстуална информация, пасажи, цели документи и т.н.

В първата част на главата е изследван проблемът за избора на най-подходящ отговор от списък с кандидати, т.е. отговор на въпрос с множествен избор. За да се избере най-добрият кандидат, системата се основава на подход с два етапа. Първо, извличане на контекстуални пасажи чрез заявка формулирана като комбинация между въпроса и всеки от кандидатите, като след това се предсказва най-вероятният отговор въз основа на извлечения доказателствен текст. Въпреки това рядко отговорът на въпроса се съдържа директно в пасажите и следователно моделите трябва да го извлекат чрез разсъждения отвъд просто съпоставяне на думи.

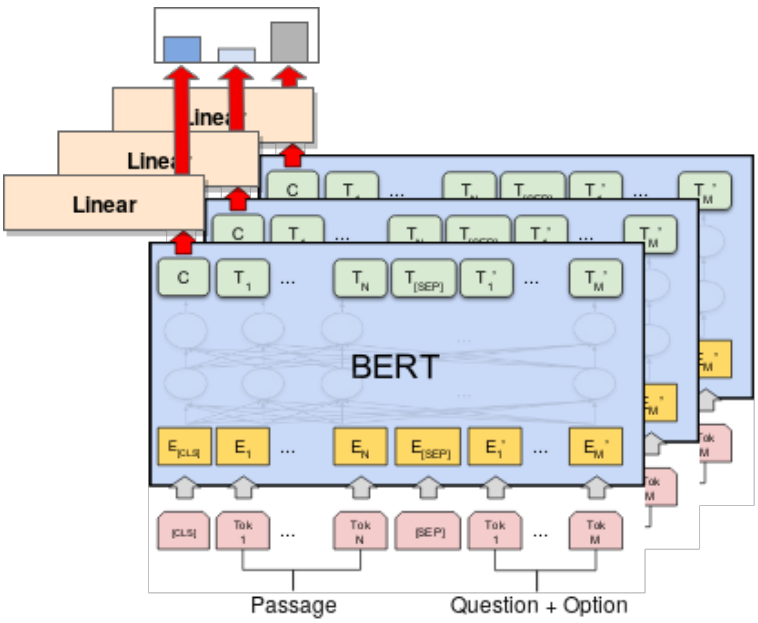
Независимо от това, едно изречение не винаги е достатъчно, за да се отговори на въпроса на клиента. Това е особено забележимо ако той се нуждае от инструкции стъпка по стъпка, за да постигне крайната си цел. Във втората

част на главата е предложена нова методология за извличане на предварително написани документи/статии, свързани с твърдения, направени в разговори в социалната платформа Twitter. В областта на разговорните агенти това може да се разглежда като преобразуване на изхода на чатбота, който обикновено е кратко изречение, в отговор в дълга форма. Този отговор може също да служи като обяснение на процес или инструкции стъпка по стъпка. В тази глава проблемът е формулирам по следния начин: полученият отговор се очаква да бъде извлечена статия за проверка на факти и следователно задачата може да се дефинира като *намиране на предишни проверени твърдения*. Изследвани са трите основни посоки, свързани с гореспоменатия проблем в тази глава, а именно: (*i*) недостига на данни – съществуващите корпуси съдържат малко обучителни примери (помалко от няколко хиляди), (*ii*) намирането на отрицателни примери – в корпуса са налични само коректно свързани двойки статия-претенции (позитивни примери) и следователно няма примери от *негативния* клас и (*iii*) учене от шумни (етикетирани с дистанционно наблюдение (*distant supervision*)) примери.

Тази глава се основава на Hardalov et al. (2019a) и Hardalov et al. (2022).

4.1 Извличане на знания

В този раздел е изследван трансфера на знания от език с много ресурси, т.е. английски, към език с малко ресурси, т.е. български, в контекста на четенето с разбиране за отговаряне въпроси с множествен избор. Повечето предишни изследвания (Pan et al., 2019; Radford et al., 2018; Tay et al., 2018; Sun et al., 2019) са едноезични и контекстите за всеки въпрос са налични предварително. Тук задачата е усложнена, като се изследват възможностите на невронен модел, използващ външно знание за разбиране в многоезична постановка. Предложеният подход се основава на многоезичен BERT (Devlin et al., 2019), допълнително обучен върху корпуса RACE (Lai et al., 2017), който съдържа над 87 000 въпроса на училищно ниво с множествен избор на английски език. За оценката на моделите е събран нов корпус за български език. Освен това са проведени експериментирано



Фигура 4.1: BERT за отговаряне на въпроси с множествен избор.

с различни стратегии за извличане на документи, и с многоезичен модел, предварително обучен върху стратифициран славянски корпус от български, чешки и полски статии в Уикипедия, и новини на руски език. Последно, трябва да се обърне внимание на недостига на ресурси в езиците и липсата на контексти на въпроси в новия корпус. Затова за всеки въпрос са извлечени релевантни пасажи от статии в Уикипедия.

4.1.1 Модел

Моделът има три компонента (вижте Фигура 4.1): (*i*) модул за извличане на контекст, който се опитва да намери подходящи обяснителни пасажи за всяка двойка въпрос-отговор от корпуса от документи (не са на английски език) (*ii*) модул за четене с разбиране на въпроси с множествен избор, предварително обучен на английски данни и след това приложен към целевия език без допълнително обучение (*zero-shot*) и (*iii*) механизъм за гласуване (*voting*), който комбинира множество пасажи от (*i*) и техните оценки от (*ii*), за да получи един (най-вероятен) отговор на въпрос.

За търсенето на подходящи пасажи за въпроси, които не са на английски език, е създаден обърнат индекс от статии в Уикипедия с помощта на Elasticsearch.

Следвайки нотацията на Devlin et al. (2019), входната последователност може да бъде представена по следния начин:

[CLS] Пасаж [SEP] Въпрос + Отговор [SEP]

Резултатите след проекционния слой са нормализирани чрез софтмакс функция. По време на допълнителното обучение параметрите на модела се оптимизират, като се максимизира логаритмичната вероятност за правилния отговор.

Намирането на доказателствени пасажи, които съдържат информация за правилния отговор, е от решаващо значение за системите за четене с разбиране. Методът за извличане на контекст може да бъде изключително чувствителен към формулирането на въпроса. Затова, в изчисляването на оценката трябва да бъдат включени и по-ниско класираните документи. вместо да се използва само най-релевантния документ. В представените експерименти е използвана проста стратегия, базирана на сумиране. Всеки резултат, върнат от инструмента за извличане на контекст, е оценен спрямо въпроса и възможните опции, като по този начин се получава списък с вероятности.

(**Биология**) Дебелата козина на бозайниците през зимата е пример за:
А. физиологична адаптация
Б. поведенческа адаптация
С. генетична адаптация
D. морфологична адаптация

(**Философия**) Според релативизма в етиката:
А. има само един морален закон, който е валиден за всички
Б. няма абсолютно добро и зло
В. хората са зли по природа
Г. има само добро, а злото е привидно

Таблица 4.1: Примерни въпроси от новия български корпус, по един за предмет. Верният отговор е маркиран в зелено.

Домейн	#Двойки	#Отговори	Дължина Въпрос	Дължина Отговор	Размер на речника
12th Grade Matriculation Exam					
Биология	437	4	10.4	2.6	2,414 (12,922)
Философия	630	4	8.9	2.9	3,636 (20,392)
География	612	4	12.8	2.5	3,239 (17,668)
История	542	4	23.7	3.6	5,466 (20,456)
Онлайн тестове по история					
Българска история	229.0	4	14.0	2.8	2,287 (10,620)
PzHistory	183	3	38.9	2.4	1,261 (7,518)
Общо	2,633	3.9	15.7	2.9	13,329 (56,104)
RACE Train - Mid and High School					
RACE-M	25,421	4.0	9.0	3.9	32,811
RACE-H	62,445	4.0	10.4	5.8	125,120
Общо	87,866	4.0	10.0	5.3	136,629

Таблица 4.2: Статистика за новия български корпус сравнен с RACE корпуса.

4.1.2 Данни

Целта е да се създаде нова задача за език с малко ресурси, като българския. Тя трябва да е възможно най-близка до постановката за четене с разбиране на тестови въпроси за езици с множество ресурси, като английския. Целта е да се оценят възможностите на обучението с трансфер в многоезична среда. Един от най-големите корпуси за тази задача (на английски език) е RACE (Lai et al., 2017), с общо 87 866 обучителни въпроса всеки с по четири възможности за отговор. Тези въпроси са зададени към 25 137 описателни контекста.

За да се изпълни тази целта е събран нов корпус за български език от 2 633 въпроса с избираем отговор, без контекст, от различни предмети: биология (16,6%), философия (23,93%), география (23,24%) и история (36,23%). В Таблица 4.1 са показани примерени въпроси от различните категории с техните възможни отговори. Верните отговори са отбелязани в зелено, а категорията на въпроса – с удебелен шрифт.

Таблица 4.2 показва разпределението на въпросите по тематична категория, дължината (в думи) както на въпросите, така и на опциите (отговорите) и размера на речника, измерен като броя на уникалните думи. Първата част на таблицата представя статистически данни за българския корпус, докато втората част е сравнение с RACE.

И накрая е разгледан размерът на речника на двата корпуса. Общият брой на уникалните думи е показан в последната колона на Таблица 4.2 (Размер на речника). За новия корпус са показани две числа на ред: първото е статистика, базирана само на двойките въпрос-отговор, докато второто, оградено в скоби, е размера на речника, включително с извлечените пасажи. Това (последното) число е по-скоро приблизително, отколкото конкретна стойност, тъй като горната му граница е броят на думите в Уикипедия и може да варира при различните стратегии за извличане на пасажи.

4.1.3 Експерименти и оценка на моделите

Допълнително обучение на BERT

Допълнителното обучение е разделено на две групи модели (*i*) Multilingual BERT и (*ii*) Slavic BERT.[1] Таблица 4.3 по-долу представя резултатите в задачата за разбиране на въпроси с множествен избор върху валидационното множество на RACE.

[1] http://github.com/deepmipt/Slavic-BERT-NER

#Епоха	RACE-M	RACE-H	Обща
Multilingual BERT			
1	64.21	53.66	56.73
2	68.80	57.58	60.84
3	69.15	58.43	61.55
Slavic BERT			
2	53.55	44.48	47.12
3	57.38	46.88	49.94

Таблица 4.3: Точност измерена върху валидационното множество на RACE след всяка епоха на обучение.

Извличане и индексиране на статии от Уикипедия

За целта е използван българският дъмп (*dump*) на Уикипедия от 2019-04-20, с общо 251 507 статии. Индексирано е заглавието и основният текст на всяка статия като обикновен текст, който ще наричаме *пасаж*. Освен това всяко поле е допълнително обработено чрез: (*i*) *ngram*: базирано на комбинации от последователни 1–3 думи (*n-gram*); (*ii*) *bg*: преобразуване на текста в малки букви, премахване на стоп думи и стеминг; (*iii*) *none*: индекс тип торба с думи (*bag-of-words*).

В експериментите е използвано подмножество от четири полета от всички възможни комбинации от анализатори (*field-analyzers*), а именно *title.bg, passage, passage.bg*, и *passage.ngram*. Приложен е анализизатор на български само върху полето *title*, тъй като има то е кратко и описателно, следователно е много чувствително към шум от стоп-думи. За разлика от това, въпросите са съставени предимно от стоп-думи, напр. *какво*, *къде*, *кога*, *как*.

За индексиране на статиите от Уикипедия са използвани две стратегии: плъзгащ се прозорец и параграф. В стратегията, базирана на прозорец, се дефинират два типа разцепване на текстовете: малки, съдържащи 80-100 думи, и големи, от около 300 думи. Накрая е използван списък с топ-N документа за всеки от възможните отговори.

Резултати от експерименти

Предварително обучение на английски за език за МЧР с множествен избор. Таблица 4.3 представя промяната на точността на моделите върху оригиналната задача за разбиране на английски, в зависимост от броя на тренировъчните епохи. В таблицата "BERT" се отнася за многоезичния модел BERT, докато "Slavic" означава BERT с предварително обучение на славянски език. Моделите са обучени допълнително върху корпуса RACE, а точността е избрана като метрика следвайки нотацията от Lai et al. (2017). Трябва да се има предвид, че въпросите в RACE-H са по-сложни от тези в RACE-M. Последната колона в таблицата, *Обща*, показва точността, изчислена за всички въпроси във валидационното множество на RACE. Ясно се вижда положителна корелация между броя на епохите и точността на модела. Освен това виждаме, че славянският BERT се представя много по-зле както на RACE-M, така и на RACE-H, което предполага, че промяната на теглата на модела към славянските езици е довела до катастрофално забравяне на научения английски синтаксис и семантика.

Постановка	**Точност**
Случаен отговор	24.89
Обучение за 3 епохи	–
+ window & title.bg & pass.ngram	29.62
+ passage.bg & passage	39.35
– title.bg	39.69
+ passage.bg^2	40.26
+ title.bg^2	40.30
+ по-голям прозорец	36.54
+ разделяне на параграфи	42.23
+ Предварително обучение със славянски езици	33.27
Обучение за 1 епоха (най-добър модел)	40.26
Обучение за 2 епоха (най-добър модел)	41.89

Таблица 4.4: Точност на българското тестово множество: изследване на поведението на модела чрез последователно добавяне/премахване на различни компоненти на модела.

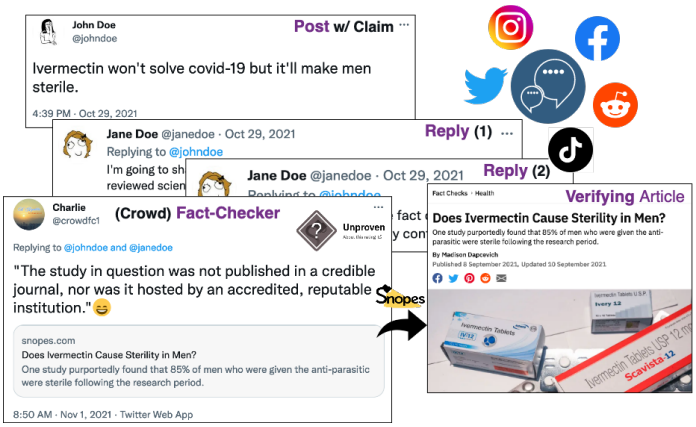
Трансфер без допълнително обучение. Тук се оценява ефективността на модела, когато той се прилага на български език за четене с разбиране на въпроси с избираем отговор. Таблица 4.4 представя изследване на значимостта на различните компоненти. Всеки ред обозначава типа на модела и добавянето (+) или премахването (–) на характеристика от постановката, представена на предишния ред. Първият ред показва ефективността на модел, който избира с равна вероятност отговор на даден въпрос от списъкът с кандидати. Следващите редове показват резултатите от експерименти, проведени с модел, обучен за три епохи на RACE.

Експериментите показват, че най-добрата комбинация от полета за заявка е *title.bulgarian^2, passage.ngram, passage, passage.bulgarian^2*, където полето *title* има незначителен принос и може да бъде премахнато с цел улеснение на изчисленията и съхраняване на пасажите. Определянето на най-добрите полета за заявка позволява да се оценят други стратегии за индексиране, т.е. по-голям прозорец (размер 1600, стъпка 400), който постига точност от 36,54%, и разделяне на параграфи, с което се постига най-високата точност от 42,23%. Това е подобрение от почти 2,0% спрямо малкия плъзгащ се прозорец и 5,7% спрямо големия.

След това е разгледано въздействието на славянския BERT. Изненадващо, той отбелязва 9% абсолютен спад в точността в сравнение с многоезичния BERT. Това навежда на мисълта, че последният вече има достатъчно знания за българския език и следователно не се нуждае от по-нататъшно адаптиране към славянските езици.

Освен това е изследвано влиянието на броя на епохите, използвани за допълнително обучение, върху производителността на модела. Наблюдава се повишаване на точността с нарастването на броя на епохите, което е в съответствие с предишните докладвани резултати за задачи на английски език. Въпреки че тази корелация не е толкова силна, колкото при оригиналната RACE задача (вижте Таблица 4.3 за сравнение), все още се наблюдава 1,6% и 0,34% абсолютно увеличение на точността, съответно за епохи 2 и 3, в сравнение с епоха 1.

Оценено е и влиянието на размера на списъка с резултати, върнат от модела за извличане на контексти, върху точността за различните категории. Допълнително е анализирана средната точност за даден размер на заявката S_q за всички



Фигура 4.2: Разговор в Twitter, проверен от тълпата. Първият туит (**публикация с твърдение**) прави твърдението, че *ивермектинът причинява стерилитет при мъжете*, след това то получава **отговори**. **Човек от тълпата (проверяващ фактите)** отговаря с връзка към **проверяваща статия** от уебсайт за проверка на факти. Сдвоява се *статията* с *туита, който прави това твърдение* (само първата публикация ✓), тъй като е неуместна (✗) за другите *отговори*.

проведени експерименти, където $S_q \in \{1, 2, 5, 10, 20\}$. Тези експерименти показват, че по-дългите списъци с резултати на заявката (т.е. съдържащи повече от 10 резултата) на опция за отговор влошават точността за всички категории, с изключение на *биология*, където се вижда малък пик при дължина 10, но въпреки това най-добрите резултати за тази категория са постигнати за списък с резултати с дължина 5. Единичен добре оформен максимум при дължина 2 е видим за *история* и *философия*. Тъй като тези две категории са най-големите, не е изненадващо, че се постига най-добра общата точност при еднакъв брой заявки.

4.2 Извличане на отговор от колекция с обяснения

В този раздел е изследван следният проблем за откриването на вече проверени твърдения: *При даден потребителски коментар трябва да се установи дали твърдението, което той прави, е било проверено преди това, като се използва колекция от вече проверени твърдения и съответните им статии* (вижте Таблица 4.5). Тази задача е неразделна част от цялостна проверка на факти (Hassan et al., 2017), а също и важна задача сама по себе си, тъй като хората често повтарят едно и също твърдение (Barrón-Cedeno et al., 2020; Vo and Lee, 2020; Shaar et al., 2021). Изследванията на този проблем са ограничени от недостига на данни, като корпусите обикновено имат около 1 000 двойки туийт–проверяваща статия (Barrón-Cedeno et al., 2020; Shaar et al., 2020, 2021), с изключение на Vo and Lee (2020), който съдържа 19K твърдения за изображения, съпоставени с 3K статии за проверка на факти.

Тази липса на големи корпуси може да бъде запълнена чрез използване на знанието на тълпата за проверка на факти. Възможно е да бъдат извлечени нови двойки статии за проверка на туитове, които след това се етикетират (ако двойката е правилно съвпадаща) автоматично с помощта на дистанционно наблюдение. Пример за такава двойка е показан на Фигура 4.2.

User Post w/ Claim: Sen. Mitch McConnell: "As recently as October, now-President Biden said you can't legislate by executive action unless you are a dictator. Well, in one week, he signed more than 30 unilateral actions." [URL] — Forbes (@Forbes) January 28, 2021		
Verified Claims and their Corresponding Articles		
(1)	When he was still a candidate for the presidency in October 2020, U.S. President Joe Biden said, "You can't legislate by executive order unless you're a dictator." http://snopes.com/fact-check/biden-executive-order-dictator/	✓
(2)	U.S. Sen. Mitch McConnell said he would not participate in 2020 election debates that include female moderators. http://snopes.com/fact-check/mitch-mcconnell-debate-female/	✗

Таблица 4.5: Илюстративни примери за задачата за откриване на предишни проверени твърдения. **Публикацията съдържа твърдение** (свързано с *законодателство и диктатура*), **Потвърдените твърдения** са част от колекция от предишни проверки на факти. В ред (*1*) провереният факт е правилно съвпадение на твърдението, направено в туита (✓), докато в (*2*) текстът също обсъжда *Sen. Mitch McConnell*, но това е различно твърдение (✗) и следователно образува неправилна двойка.

4.2.1 Новосъздаден корпус: CrowdChecked

Събиране на данни

Поради популярността на уебсайта Snopes както сред интернет потребителите, така и сред изследователите (Popat, Kashyap and Mukherjee, Subhabrata and Strötgen, Jannik and Weikum, Gerhard, 2016; Hanselowski et al., 2019; Augenstein et al., 2019; Tchechmedjiev et al., 2019), той е използван като основен източник за проверка на факти. Източник за събиране на потребителски съобщения е Twitter. Тези съобщения могат да съдържат твърдения и проверки на факти свързани с тези твърдения.

Постановката за събиране на данни е подобна на тази в Vo and Lee (2019). Първо се дефинира заявка за избиране на туитове, които съдържат връзка към статия с проверка на факти от Snopes (*url:snopes.com/fact-check/*). Туитът е или отговор (*reply*), или цитат (*quote*), а не ретуит (*retweet*).[2] Примерен резултат от заявката е показан на Фигура 4.2, където туитът написан *от човека от тълпата проверяващ фактите* съдържа връзка към такава статия. След това уместността на твърдението (ако има такова), направено в първия туит (корена на разговора) и последния отговор се оценява, за да се извлекат нови двойки туит-проверена статия.

По този начин са събирани всички туитове от октомври 2017 г. до октомври 2021 г.. В резултат на това са получени общо 482 736 уникални двойки. Допълнително са събирани 148 503 туита с отговори и 204 250 туита от началото на разговора (корена).[3]

[2]Изключват се ретуитове, тъй като те не съдържат коментари, а по-скоро споделят предишни туитове.

[3]Сборът от уникалните отговори и туитовете от началото на разговори не е равен на броя туитове за проверка на фактите, тъй като повече от един туит може да отговори на един и същ коментар.

Корпус	**Туитове**‡	**Думи**			**Речник**
	|Уникални|	Средно	50%	Макс.	|Уникални|
CrowdChecked (Нов)	316,564	12.2	11	60	114,727
CheckThat '21	1,399	17.5	16	62	9,007

Таблица 4.6: Статистика за новия корпус спрямо CheckThat '21. ‡Броят на уникалните туитове е по-малък в сравнение с общия брой двойки туит-статия, тъй като един туит може да бъде проверен от множество статии.

Събиране на туитове (структура на разговора)

Важно е да се отбележи, че туитът съдържащ *'проверка на фактите'* може да бъде част от разговор с множество реплики, следователно коментарът, на който се отговаря (предишна реплика), не винаги изразява твърдение, което текущият туит цели да провери. За да разберем по-добре това явление, е направен ръчен анализ на разговорите. Разговорите са организирани по подобен начин, показан на Фигура 4.2, т.е. коренът е първият коментар, след което може да има дълга дискусия, последвана от коментар, който проверява фактите (този със Snopes връзката).

Сравнение със съществуващи корпуси

Тук е сравнен новият корпус с подобен такъв от състезанието CLEF-2021 CheckThat '21 за откриване на предишни проверени твърдения в туитове (Shaar et al., 2021) (наричан *CheckThat '21* в останалата част от главата). Съществуват други подобни корпуси, но те съдържат по-малко примери (Barrón-Cedeno et al., 2020), от различен домейн (Shaar et al., 2021), не на английски език (Elsayed et al., 2019) или са мултимодални (Vo and Lee, 2020).

Таблица 4.6 представя сравнение между *CrowdChecked* и *CheckThat '21* по отношение на броя на примерите, дължината на туитовете и размера на речника. Преди да се изчислят тези статистики, текстът е конвертиран в малки букви и са премахнати всички URL адреси, манипулатори на Twitter, английски стоп думи и пунктуация. Можем да видим, че *CrowdChecked* съдържа два порядъка повече примери, малко по-кратки туитове (но максималната дължина остава приблизително същата, което може да се обясни с ограничението на думите в Twitter) и размер на речника, който е в порядък по-голям. Обаче много примери в *CrowdChecked* са неправилни съвпадения и затова е използвано дистанционно наблюдение за тяхното етикетиране. Новосъздадените корпуси с правилни съвпадения са показани в Таблица 4.7. Трябва да се подчертае, че няма абсолютно никакво припокриване между CrowdChecked и CheckThat '21 по отношение на туитове/твърдения.

Етикетиране на данни (дистанционно наблюдение)

Данните са етикетирани чрез два подхода за дистанционен надзор: (*i*) базиран на приликата на Жакард (*Jaccard*) между туита и неговата проверяваща статия и (*ii*) базиран на предсказанията на модел, обучен на CheckThat '21.

За да се оцени качеството на получените етикети, е направена ръчна анотация, целяща да изчисля броя на *коректни двойки* (т.е. туит–статия двойки, където статията проверява твърдението в туита). Предишните наблюдения върху данните подсказват, че случайното избиране на примери от корпуса с туитове

не е подходящо, защото той включва предимно двойки, които имат много малко припокриващи се думи, което често е индикатор, че текстовете нямат връзка помежду си.

4.2.2 Метод

Общата схема Като базов модел се използва Sentence-BERT (SBERT). Запазена е основната архитектура, предложена от Reimers and Gurevych (2019), но са използвани допълнителни характеристики, трикове при обучение и функции на грешката. Всички те са описани в следващите раздели. Входът на модела е двойка от туит и статия за проверка на фактите, които са кодирани, както следва:

- Потребителски туит: [CLS] *Текст на туита* [SEP]
- Проверяваща статия: [CLS] *Заглавие* [SEP] *Подзагавие* [SEP] *Проверено твърдение* [SEP]

Обучаваните модели използват грешка, базирана на ранкиране с множество отрицателни примери (multiple negatives ranking) (MNR) (Henderson et al., 2017) (вижте Формула 4.1). В допълнение е предложен нов вариант на MNR грешката, който отчита шума в корпуса.

Обогатена схема Тук се използва методът, предложен в най-добре представилата се система от състезанието CheckThat '21 (Chernyavskiy et al., 2021). Той се състои от независими компоненти за оценка на лексикални (TF.IDF) и семантични (SBERT) сходства. Моделите SBERT използват същата архитектура и входен формат, както е описано в *'Общата схема'* по-горе. Въпреки това Chernyavskiy et al. (2021) използва множество от модели.

Добавен е и температурен параметър (τ) в MNR функцията на грешката. За да се стабилизира процеса на обучение, този параметър се оптимизира, както е предложено в Chernyavskiy et al. (2022).

Обучение с шумни данни

Предложен е нов метод, базиран на самоадаптивно обучение (Huang et al., 2020), за да се отчете възможният шум в етикетираните данни чрез дистанционно наблюдение. Този метод е предложен за класификационни задачи с функция на грешката кръстосана ентропия; обаче той трябва да бъде модифициран, за да се използва със MNR. За целта се прилага итеративно обновяване на етикетите y на база на предсказанията на текущия модел, започвайки след избрана епоха, която е хиперпараметър:

$y^r \leftarrow \alpha \cdot y^r + (1-\alpha) \cdot \hat{y}$,

където y^r е текущият обновен етикет (първоначално $y_r = y$), $\hat{y}$ е предсказанието на модела, а α е хиперпараметър на импулса (*momentum*) ($\alpha = 0{,}9$).

Адаптираната версия на MNR функцията на грешката се дефинира по следния начин:

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m} {y^r}_i \Big(\frac{c_i^T v_i}{\tau} - \log \sum_{j=1}^{m} \exp(\frac{c_i^T v_j}{\tau}) \Big) \tag{4.1}$$

Ако зададем $y_i^r = 1$, тогава Формула 4.1 става еквивалентна на дефиницията за MNR. Параметърът τ е температурата.

В предложения от Huang et al. (2020) за самоадаптивно обучение се въвеждат тегла $w_i = \max_{j \in \{1,..,L\}} t_{i,j}$, където t_i е коригираният позиционно кодиран (*one-hot encoded*) целеви вектор в задача за класификация с L класове. След

прилагане на двете модификации, въздействието на всеки пример за обучение е пропорционално на квадрата на коригирания етикет, т.е. във Формула 4.1 y_i^r вече е на квадрат.

Пренареждане на резултатите

Използвана е процедурата за пренареждане на резултатите, предложена от Chernyavskiy et al. (2021). Те използват LambdaMART (Wu et al., 2010) модел. Входящите данни са реципрочните рангове (позиция в класирания списък с твърдения) и предсказаните резултати за релевантност (2 фактора) въз основа на близостта изчислена от моделите TF.IDF и SBERT (2 модела) между туита и твърдението, твърдението+заглавието и твърдението+заглавието+подзаглавието (3 комбинации) – това прави общо дванадесет характеристики в ансамбъла и четири в единичния модел.

4.2.3 Експерименти

Корпуси

Таблица 4.7 показва статистика за размерите на множествата в CrowdChecked и CheckThat '21. Тези множества се използват във всички експерименти, макар и понякога смесени заедно.

Първата група (CrowdChecked) показва множестото, етикетирано чрез дистанционно наблюдение. Тъй като положителните двойки са анотирани с дистанционно наблюдение, а не от хора, те са включени като част от обучителното множество. Всяко показано множество се получава с помощта на различна мярка за сходство (Жакард или косинус) и съответно праг. От общия брой 332K събрани двойки туит–статия в CrowdChecked, са филтрирани подгрупи от размери между 3,5K и 49K примери.

Втората група описва корпуса CheckThat '21. При него са запазени оригиналните обучително, валидационно и тестово множества. Във всеки от представените експерименти моделите се валидират и тестват върху съответните множества от CheckThat '21, докато данните за обучение могат да бъдат смесени с CrowdChecked.

Корпус	**Множество**	**Праг**	**Туит-Статия Двойки**
CrowdChecked (новият корпус)	Train	-	332,660
	Train *Jaccard*	0.30	27,387
		0.40	12,555
		0.50	4,953
	Train *Cosine*	0.50	48,845
		0.60	26,588
		0.70	11,734
		0.80	3,496
CheckThat '21	Train	-	999
	Валидационно	-	199
	Тестово	-	202

Таблица 4.7: Статистика за събраните корпуси по отношение на двойките туит-статии с проверка на фактите.

Модел	MRR	P@1	MAP@5
Базови модели (CheckThat '21)			
Извличане на документи (Shaar et al., 2021)	76.1	70.3	74.9
SBERT (CheckThat '21)	79.96	74.59	79.20
CrowdChecked (Нов корпус)			
SBERT (jac $>$ 0.30)	81.50	76.40	80.84
SBERT (cos $>$ 0.50)	81.58	75.91	81.05
(Pre-train) CrowdChecked, (Fine-tune) CheckThat '21			
SBERT (jac $>$ 0.30, Seq)	**83.76**	**78.88**	**83.11**
SBERT (cos $>$ 0.50, Seq)	82.26	77.06	81.41
(Mix) CrowdChecked и CheckThat '21			
SBERT (jac $>$ 0.30, Mix)	83.04	78.55	82.30
SBERT (cos $>$ 0.50, Mix)	82.12	76.57	81.38

Таблица 4.8: Оценка върху тестовите данни на CheckThat '21. В скоби е името на обучаващото множество, т.е. стратегия за избор на *Jac*card или *Cos*ine, при *(Seq)* моделът първо се обучава върху CrowdChecked и след това върху CheckThat '21, при *(Mix)* – данните от двата корпуса се смесват. Най-високите резултати са с **удебелен** шрифт.

Резултати от експерименти

Анализ на стойностите на прага Таблица 4.8 показва резултатите, групирани въз основа на използваните данни за обучение. Във всяка група се включват двата модела с най-добри резултати.

Виждаме, че всички модели базирани на SBERT постигат по-високи резултати с 4–8 абсолютни точки MAP@5 повече спрямо базовия модел за извличане на документи. Интересното е, че обучението само върху дистанционно контролирани данни е достатъчно, за да получим по-висок с повече от 1,5 MAP@5 точки резултат от този на SBERT, обучен на CheckThat '21. Освен това ефективността на двете стратегии за етикетиране на данни (т.е. Jaccard и Cosine) е относително близка, което предполага сравнимо количество шум в двата корпуса.

Добавянето на повече, етикетирани с дистанционно наблюдение, данни е от полза за модела независимо от стратегията. Единственото изключение е спадът в метриките, когато прагът на Жакардовото сходството се намали от 0,5 на 0,4.

Моделиране на шумни данни Изследвани са и ефектите от предложените промени в подхода за обучение на SBERT: (*i*) разместване (*shuffling*) и обучаем температурен параметър (*temperature*), (*ii*) модификация на функцията на грешката MNR за самоадаптивно обучение (*self-adaptive*) с допълнителни тегла. За това е използван е моделът *(jac $>$ 0,30, mix)*, тъй като постигна най-висок резултат. В Таблица 4.9 е изследвана всяка от тези модификации, като са те са добавяни итеративно към базовия SBERT модел.

4.3 Обобщение

В тази глава бяха представени два подхода за извличане на отговори от външни източници на знания, а именно: (*i*) трансфер (без допълнително обучение) към език с малко ресурси за задачата за отговаряне на въпроси с множествен избор, базиран на набор от извлечени контексти с доказателствени параграфи

Model	**MAP@5**	
	Валидцаионно	Тестово
DIPS (Mihaylova et al., 2021)	93.6	78.7
NLytics (Pritzkau, 2021)	-	79.9
Aschern (Chernyavskiy et al., 2021)	94.2	88.2
SBERT (jac $>$ 0.30, Mix)	90.0	82.3
+ shuffling & trainable temp.	92.4	82.6
+ self-adaptive training (Eq. 4.1)	92.6	83.6
+ loss weights	*92.7*	*84.3*
+ TF.IDF + Re-ranking	93.1	89.7
+ TF.IDF + Re-ranking (ens.)	**94.8**	**90.3**

Таблица 4.9: Резултати на CheckThat '21 (валидационно и тестово множество). Сравнява се предложения модел и неговите компоненти (добавени последователно) с най-съвременните модели в областта. Най-добрите резултати са с **удебелен** шрифт.

от външна база от знания и (*ii*) извличане на отговор от колекция с обяснения, т.е. предварително написани дълги отговори (тип документи или статии).

Първо беше изследвана задачата за избор на отговор при четене с разбиране за езици с малко ресурси. Беше събран нов корпус с 2 633 въпроса на български език без пояснителен контекст от матури след дванадесети клас по география, философия, история и биология и онлайн тестове по история. По-точно беше проектиран цялостен подход базиран на многоезичен BERT модел (Devlin et al., 2019), който е предварително обучен върху голям корпус за четене с разбиране на английски език и източник на знания от множество домейни (Уикипедия). Резултатите от експериментите показаха, че допълнителното предварително обучение на английския RACE корпус помага на моделите, докато предварителното обучение върху данни от славянски езици вреди. Тази разлика най-вероятно се дължи на катастрофално забравяне. Разделянето на абзаци, n-грами, премахването на стоп-думи и стеминга допълнително помагат на модела за извличане на документи да намери по-добри доказателствени пасажи. Така моделът може да постигне точност от 42.23%, което е значително по-високо от резултата на базовите модели – 24,89% и 29,62%.

След това беше преставан CrowdChecked– нов кропус за откриване на предишни проверени твърдения с повече от 330 000 двойки туитове и съответните статии с проверени факти, публикувани от хора от тълпата, проверяващи твърдения. Допълнително бяха проучени две техники за етикетиране на двойки туит-статия с помощта на дистанционно наблюдение, базирано на сходството на Жакард и предсказанията на многослойна невронна мрежа. По този начин са създадени нови множества за обучение от 3,5K–50K примера. Предложен беше подход за обучение с шумни данни, използвайки самоадаптивно обучение и допълнителни тегла във функцията на грешката. Освен това беше потвърдена ефективността на новите данни, които водят до значителни подобрения от четири процента спрямо съвременни базови модели по отношение на метриките MRR, P@1 и MAP@5. Трябва да се отбележи, че всички базови модели са обучени върху ръчно анотирани данни (Shaar et al., 2021). И накрая беше постигнат по-добър, с две точки върху корпуса CheckThat '21, резултат спрямо текущите най-съвременни модели. Моделът постигна MAP@5 от 90.3, когато се използва събраният нов корпус и предложеният метод.

Глава 5. Усъвършенствани методи за разговор

Тази глава изследва усъвършенствани методи за разговор, които надхвърлят един език и индивидуални модели. Първо са обсъдени цялостни (*end-to-end*) генеративни модели. За разлика от моделите, обсъдени в предишните глави, генеративните агенти могат да управлява диалога и да предлагат нови отговори, които не са били срещани досега в разговор, без да зависят от външни източници или РЕЕ компоненти.

Допълнително е предложен нов подход за избиране на следващото изказване в разговора от множество кандидати, получени от много източници, например генерирани с помощта на последователност към последователност модели или извлечени от база от знания. Предложените подходи са оценени, използвайки голям корпус, събран от реални разговори с цел поддръжка на клиенти в социалните медии (Twitter) между компании и техните потребители.

И накрая са изследвани по-комплексни методи от едноезично обучение и трансфер без допълнително обучение. Представен е нов корпус за отговор на въпроси с множествен избор, обхващащ шестнадесет езика от осем езикови семейства. В допълнение този корпус е използван, за да се оценят възможностите на най-съвременните многоезични модели за междуезичен трансфер. Този раздел развива и разширява някои от идеите, представени в Глава 4, *Извличане на знания*.

Тази глава се основава главно на Hardalov et al. (2018), Hardalov et al. (2019b) и Hardalov et al. (2020b).

5.1 Корпус от разговори на тема обслужване на клиенти

Като цяло данните и ресурсите, които биха могли да се използват за обучение на чатбот за обслужване на клиенти, са много оскъдни, тъй като компаниите пазят разговорите на собствени, недостъпни за широката общественост, системи. Това се дължи от една страна на поверителността на клиентските данни, и от друга на това, че компаниите не искат да оповестяват публично своето ноу-хау и често срещаните проблеми относно своите продукти и услуги.

Тази ситуация се промени с публикуването на нов корпус в платформата Kaggle, наречен *Customer Support on Twitter*.[1] Той представлява голяма колекция от скорошни туитове и отговори, която е предназначена да спомогне иновациите в разбирането на естествения език и разговорните модели. Също така цели да помага при изучаването на модерни практики за обслужване на клиенти. Корпусът съдържа над 3 милиона туита от 20 големи компании като Amazon, Apple, Uber, Delta, Spotify, и други.

[1] https://www.kaggle.com/thoughtvector/customer-support-on-twitter

Общ	
# words (in total)	26,140
Мин. # реплики в диалог	2.00
Макс. # реплики в диалог	106.00
Среден # реплики в диалог	2.6
Среден # думи във въпрос	20.00
Среден # думи във отговор	25.88
# двойки диалози	49,626
Обучаващо множество: # диалози	45,582
Тестово множество: # диалози	4,044

Таблица 5.1: Обща статистика за корпуса.

Тъй като темите за поддръжка на клиенти от различни организации обикновено не са свързани една с друга, изследванията са насочени само върху туитове, свързани с компанията Apple. Това позволява да се фокусира върху малък набор от теми, които са свързани с една компания – ситуация, по-близка до сценария от реалния свят. Таблица 5.1 показва статистики за корпуса.

5.2 Цялостни генеративни агенти

Бързото разпространение на мобилни и преносими устройства даде възможност за създаването на редица нови продукти и услуги. И все пак постави допълнително напрежение върху обслужването на клиенти, тъй като потребителите в днешни дни очакват да получават денонощно информация за техните поръчки или отговори на основни въпроси като 'Защо нямам Интернет връзка?' и 'В колко е следващият влак от София за Варна?'.

Чатботовете са особено подходящи за задачата, тъй като са автоматични: напълно или частично. Освен това те са приложими от технологична гледна точка, тъй като обикновенно областта, за която отговарят, е тясна. В резултат на това общите разговори (*chit-chat*) са сведени до минимум, а разговорните агенти служат предимно като устройства за отговаряне на въпроси. Нещо повече, възможно е да бъдат обучавани върху реални разговори. В тази глава са показани експерименти с такива разговори в Twitter и са сравнени два типа агенти: (*i*) базирани на извличане на информация (information retrieval) (Изв. Инф.) и (*ii*) базирани на невронни мрежи за отговори на въпроси. Допълнително са изследвани мерки за семантично сходство. Стандартните метрики като ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) и METEOR (Banerjee and Lavie, 2005), които произтичат от машинния превод или резюмирането на текст, не са много подходящи за чатботове.

5.2.1 Метод

Предварителна обработка

Тъй като Twitter има свои собствени специфики по отношение на дължина на текста и стил на писане, стандартното токенизиране на текст обикновено не е подходящо за туитове. Затова за предварителна обработка на данните е използван специализиран Twitter токенизатор (Manning et al., 2014). Данните са изчистени,

като са заменени съкращенията, жаргонните думи, URL адресите с <*url*>, всички споменавания на потребители с <*user*> и всички хаштагове с <*hashtag*>. За обучението на модела се използват най-честите N, всички останали са заменени със специален символ <*unk*>.

Извличане на информация

Подходът базиран на Изв. Инф. може да се дефинира по следния начин: при дадени потребителски въпрос q' и списък от двойки предварително зададени въпроси и техните отговори $(Q, A) = \{(q_j, a_j)|j = 1, \ldots, n\}$, да се намери най-сходният въпрос q_i в корпуса за обучение, който потребителят е задавал преди това, и да се върне отговор a_i, който операторът за клиентска поддръжка е дал на q_i. Сходството между q' и q_i може да се изчисли по различни начини, но най-често това се прави с помощта на косинусово сходство между съответните TF.IDF-претеглени вектори.

Поредица в последователност

Енкодерът използва двупосочна рекурентна невронна мрежа РНМ, базирана на LSTM (Hochreiter and Schmidhuber, 1997). Той кодира входната последователност $x = (x_1, \ldots, x_n)$ и изчислява последователност от скрити състояния $(\overrightarrow{h_1}, \ldots, \overrightarrow{h_m})$, както и обратна последователност $(\overleftarrow{h_1}, \ldots, \overleftarrow{h_m})$. Декодерът е еднопосочен LSTM-базиран РНМ и предсказва изходната последователност $y = (y_1, \ldots, y_n)$. Всяко y_i се предсказва с помощта на повтарящото се състояние s_i, предишната предсказана дума y_{i-1} и контекстен вектор c_i. Последният се изчислява с помощта на механизъм за внимание като претеглена сума върху изхода на енкодера $(\overrightarrow{h_j}, \overleftarrow{h_j})$, както е предложено от Bahdanau et al. (2015).

Трансформатор (*Transformer*)

Моделът Трансформатор беше предложен от Vaswani et al. (2017) и показа много високи резултати при машинен превод, напр. постига най-висок резултат на данните от WMT2014 за англо-немски и англо-френски превод. Подобно на модела Seq2seq, Трансформаторът има енкодер и декодер. Енкодерът се състои от подредени идентични слоеве, базиран на самовнимание с няколко глави и проста напълно свързана по отношение на позицията невронна мрежа. Декодерът е подобен, като в допълнение към двата подслоя в енкодера, той въвежда трети подслой, който прилага внимание чрез няколко глави върху изходите на стека на енкодерите. Основното предимство на модела Transformer е, че той може да бъде обучен значително по-бързо в сравнение с рекурентни или конволюционни невронни мрежи.

5.2.2 Експерименти

Таблица 5.2 сравнява резултатите за трите модела (Изв. Инф., Seq2seq и Transformer), като се използват мерки за припокриване на думи като BLEU@2, която използва само униграми и биграми и ROUGE-L (Lin and Och, 2004), която използва най-дълга обща подпоследователност (*longest common subsequence*).

Таблица 5.3 показва резултатите за същите три системи, но с помощта на мерки за семантична оценка, а именно Embedding Average (с косинусово сходство), Greedy Matching и Vector Extrema (с косинусово сходство). И за трите мерки са използвани предварително обучените word2vec ембединги на Google, защото те

	Оценка чрез припокриване на думи	
	BLEU@2	ROUGE-L
Изв. Инф. - BM25	13.73	22.35
Seq2seq	**15.10**	**26.60**
Transformer	12.43	25.33

Таблица 5.2: Резултати от експерименти с мерки, базирани на припокриване на думи (*word overlap*).

	Семантична оценка		
	Embedding Average	Greedy Matching	Vector Extrema
Изв. Инф. - BM25	76.53	29.72	37.99
Seq2seq	**77.11**	**30.81**	**40.23**
Transformer	75.35	30.08	39.40

Таблица 5.3: Резултати, базирани на мерки за семантична оценка.

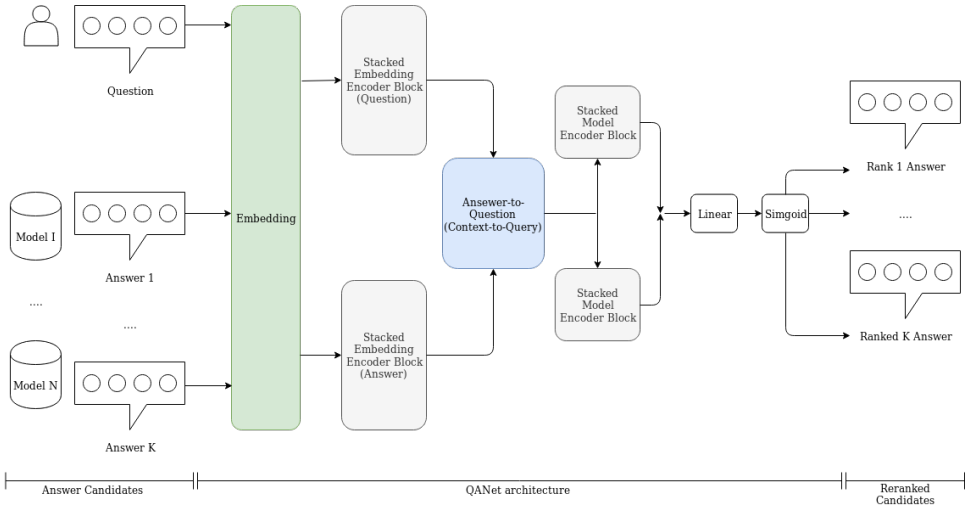
не се научават по време на обучението, което помага да се избегнат пристрастия, както беше предложено в (Liu et al., 2016; Lowe et al., 2017).

Резултатите от експериментите показват, че *Seq2seq* се представя най-добре по отношение на всичките пет мерки за оценка. За групата от семантични мерки, той превъзхожда другите системи по отношение на Embedding Average с +0.58, по отношение на Greedy Matching с +0.73 и по отношение на Vector Extrema +0.83 (абсолютни точки). Нещо повече, Seq2seq е най-добрият модел по отношение на мерките за оценка, базирани на припокриване на думи, получавайки стойности от 15.10 на BLEU@2 (+1.37 пред втория) и 26.60 на ROUGE-L (+1.27 в сравнение с втората най-добра система). Моделът *Transformer* се класира на второ място по три от мерките за оценка: Greedy Matching, Vector Extrema и ROUGE-L. И накрая, моделът за извличане на информация (*Изв. Инф.*) постига вторите най-добри резултати по отношение на BLEU@2 и Embedding Average, но е най-ниско класираният според другите три мерки за оценка. Това показва предимствата на генеративните модели, базирани на невронни мрежи, над тези базирани на извличане на информация.

5.3 Избор на отговор от множество източници

Нарастващата популярност на интелигентни устройства, лични асистенти и онлайн системи за обслужване на клиенти мотивира изследователската общност да се фокусира върху разработеното нови методологии за автоматично отговаряне на въпроси и чатботове. В областта на разговорните агенти доминират два основни вида системи: (i) базирана на извличане на информация и (ii) генеративна. Докато първите извличат ясни и консистенции отговори, вторите са гъвкави и дават възможност за генериране на нови несрещани досега изречения.

В тази дисертацията фокусът е върху намирането на най-подходящия отговор на въпрос, където всеки кандидат може да бъде произведен от различна система, например базирана на знания, базирана на правила, дълбока невронна



Фигура 5.1: Предложеният метод за преподреждане на отговорите, базиран на архитектурата QANet.

мрежа, извличане на информация и т.н. По-точно, е разработен подход за преподреждане на двойки въпрос-отговор, базирана на машинно четене с разбиране.

5.3.1 Модел за преподреждане

Предложен е подход за преподреждане на отговорите чрез класификатор, базиран на QANet (Yu et al., 2018), една от най-съвременните архитектури за машинно четене с разбиране, който да оцени дали даден отговор е подходящ за целевия въпрос. След това постериорните вероятности на класификатора се използват, за да се подредят кандидат отговорите, както е показано на Фигура 5.1.

Семплиране на негативни примери

Целта е да се разграничат "добрите" от "лошите" отговори, но оригиналният корпус съдържа само валидни, т.е. "добри" двойки въпрос-отговор. Затова се използва *семплиране на негативни примери* (Mikolov et al., 2013), при което се замества оригиналния отговор на даден въпрос с произволен отговор от обучаващото множество. Освен това, се сравнява косинусовото сходство на база думи между оригиналния и семплирания отговор и в някои редки случаи "лош" отговор се превръща в "добър", ако е твърде подобен на оригиналния "добър" отговор.

QANet Архитектура

При задачата за машинното четене с разбиране целта е да се отговори на въпрос, като се извлече низ от даден текстов контекст. Тук е използван този модел, за да се оцени уместността на дадена двойка въпрос-отговор.

Първият слой на мрежата е стандартен ембединг (*embedding*) слой, който трансформира думите в нискоразмерни плътни вектори. След това върху представянията се добавя двуслойна магистрална (*highway*) мрежа (Srivastava et al., 2015). Това позволява на мрежата да регулира информационния поток с помощта на пропусквателен механизъм.

Включени са експерименти с два типа входни ембединги. Първо, са използвани 200-измерни GloVe (Pennington et al., 2014) вектори, обучени върху 27 милиарда публикации от Twitter. След това, тяхното представяне е сравнено с

ELMo (Peters et al., 2018), наскоро предложен начин за обучение на контекстуализирани представяния на думи. В ELMo векторите за думите се получават чрез научени функции за активиране на вътрешните състояния на дълбок двупосочен езиков модел.

Ембединг слоят на енкодера се базира на конволюция, последвана от самовнимание (*self-attention*) (Vaswani et al., 2017) и многослойна невронна мрежа. Резултатът от слоя е $f(layernorm(x)) + x$, където $layernorm$ е операцията за нормализиране на слоя. Резултатът отново се преобразува в $\#words \times d$ чрез 1D конволюция. Входните и ембединг слоевете се обучават отделно за въпроса и отговора.

Слоят за внимание (*attention layer*) е стандартен модул за модели за машинно четене с разбиране. Използват се имената *answer-to-question* (*A2Q*) и *question-to-answer* (*Q2A*), които са също известни като *context-query* и *query-context*, съответно.

Слоят за внимание е последван от слой за моделиране, който приема като вход конкатенацията на $[a; a2q; a \odot a2q; a \odot q2a]$ – редове от оригиналните матрици. За изходния слой се обучават две различни представяния, които се получават, като се предава изходът от слоят за моделиране към два остатъчни (*residual*) блока, прилагайки дропаут (*dropout*) (Srivastava et al., 2014) само към входовете на първия. Предсказаната вероятност се представя като $P(a|q) = \sigma(W_o[M_0; M_1])$. Теглата се оптимизират чрез минимизиране на грешката от двоична крос ентропия (*binary cross-entropy loss*).

Избор на отговор

Експериментирано е с две стратегии за избор на отговор: (i) *max* – базирана на максимум и (ii) пропорционална извадка след софтмакс нормализация. Първата стратегия е стандартна и избира отговора с най-висок резултат, докато втората връща произволен отговор с вероятност, пропорционална на резултата, върнат от софтмакс слоя, с цел увеличаване на разнообразието на отговорите.

И за двете стратегии се използва линейна проекция, приложена към изхода на последния остатъчен блок (*residual block*) на модела, който е показан като *линеен блок* на Фигура 5.1. Това може да бъде формализирано, както следва: $o(q, a_k) = W_o[M]$, където M е конкатенацията на изходите на един или повече остатъчни блокове на модела.

Емпиричните експерименти показват, че изборът на отговор въз основа на стратегията *max* не винаги се представя добре. Може да се постигне значително подобрение, като се използва пропорционално семплиране след софтмакс нормализация, вместо винаги да се избира отговора с най-висока вероятност. В показаните експерименти Ans се моделира като случайна променлива, която следва категориално разпределение върху $K = |A|$ събития (кандидати за отговори).

5.3.2 Резултати от експерименти

Помощна задача: класифициране на уместността на двойки въпрос–отговор

Таблица 5.4 показва резултатите за помощната задача за класифициране на уместността въпрос–отговор. Първата колона е името на модела. След това са показани три колони, обозначаващи типа на ембедингите, размера на скрития слой и броя на главите в модела (*heads*). Точността на модела е описана в Последната

Модел	Тип на ембединг	d_model	Глави	Точност
Мажоритарен клас	–	–	–	50.52
QANet	GloVe	64	4	80.58
		64	8	82.83
		128	8	83.42
QANet	ELMo (на ниво токен)	64	4	82.92
		64	8	83.88
		128	8	83.48
QANet	ELMo (на ниво изречение)	64	8	84.09
		128	8	**85.45**

Таблица 5.4: Помощна задача: резултати от класификацията на уместността между въпрос–отговор.

	Припокриване на думи		Семантично сходство		
Модел	**BLEU@2**	**ROUGE_L**	**Emb Avg**	**Greedy Match**	**Vec Extr**
Transformer	12.43	25.33	75.35	30.08	39.40
Изв. Инф.-BM25	13.73	22.35	76.53	29.72	37.99
Seq2seq	15.10	26.60	77.11	30.81	40.23
QANet върху Изв. Инф. *(Индивидуален)*	14.92 ± 0.13	23.30 ± 0.12	77.47 ± 0.06	30.40 ± 0.06	39.63 ± 0.06

Таблица 5.5: Основна задача: производителност на отделните модели. Резултатите от единичните модели са показани в Таблици 5.2 и 5.3.

колона. Тъй като корпусът е балансиран (генерират се приблизително 50% положителни и 50% отрицателни примери), точността е подходяща мярка за оценка за тази задача. Горният ред на таблицата показва точността при предсказване на мажоритарния клас. Следващите редове са резултатите за пълния модел базиран на QANet при използване на различни типове ембединги. Можем да видим, че контекстуализираните ембединги на ниво изречение са за предпочитане пред използването на прости ембединги на ниво думи като GloVe или ELMo ембединги на ниво токен. Трябва да се отбележи, че ELMo на ниво токен постига по-висок резутат от GloVe само когато размерът на мрежата е малък. Когато броят на параметрите нараства ($d_{model} = 128$, $\#Heads = 8$), няма значителна разлика между двата модела.

Избор/Генериране на отговор: Индивидуални модели

Таблица 5.5 показва точността на отделните модели: извличане на информация (information retrieval) (Изв. Инф.), последователност към последователност (Seq2seq) и Transformer. Същата постановка се използва за експериментите, описани в Раздел 5.2. Таблицата е организирана по следния начин: Първата колона съдържа името на модела, използван за получаване на най-добрия отговор. Втората и третата колона отчитат мерките за припокриване на думите: (i) BLEU@2, която използва униграмни и двуграмни съвпадения между хипотезата и референтното изречение, и (ii) ROUGE-L, която използва LSC. В последните три колони са описани мерките за семантично сходство: (i) Embedding Average (Emb Avg), (ii) Greedy Matching (Greedy Match) и (iii) Vector Extrema (Vec Extr), които са изчислени чрез косинусово сходство. За целта се използват стандартни

Модел	Припокриване на думи		Семантично сходство		
	BLEU@2	**ROUGE_L**	**Emb Avg**	**Greedy Match**	**Vec Extr**
Случен отговор	14.52 ± 0.12	23.41 ± 0.12	77.21 ± 0.06	30.24 ± 0.07	38.25 ± 0.20
QANet+GloVe					
d=64, h=4	15.18	24.13	78.38	31.14	**40.85**
Softmax	15.81 ± 0.09	24.53 ± 0.05	78.32 ± 0.08	31.10 ± 0.03	40.51 ± 0.12
d=64, h=8	15.41	23.62	78.48	30.97	40.81
Softmax	15.90 ± 0.06	24.39 ± 0.03	78.38 ± 0.04	31.11 ± 0.02	40.66 ± 0.06
d = 128, h = 8	15.94	24.59	78.29	31.19	40.63
Softmax	16.04 ± 0.08	24.71 ± 0.06	78.36 ± 0.07	31.20 ± 0.07	40.70 ± 0.05
QANet+ELMo (Токен)					
d = 64, h = 4	15.23	23.48	78.25	30.77	40.22
Softmax	15.77 ± 0.15	24.44 ± 0.09	78.27 ± 0.03	31.06 ± 0.05	40.46 ± 0.11
d = 64, h = 8	15.30	23.41	**78.54**	30.97	40.19
Softmax	15.86 ± 0.07	24.40 ± 0.06	78.36 ± 0.08	31.11 ± 0.04	40.49 ± 0.05
d = 128, h = 8	15.24	23.59	78.34	30.90	40.19
Softmax	15.89 ± 0.08	24.55 ± 0.10	78.33 ± 0.06	31.11 ± 0.05	40.40 ± 0.05
QANet+ELMo (Изречение)					
d = 64, h = 8	15.48	23.88	78.44	30.96	40.33
Softmax	16.00 ± 0.14	24.50 ± 0.33	78.34 ± 0.10	31.13 ± 0.08	40.56 ± 0.09
d = 128, h = 8	15.64	24.13	78.52	31.14	40.63
Softmax	$\mathbf{16.05 \pm 0.06}$	$\mathbf{24.81 \pm 0.08}$	78.40 ± 0.07	$\mathbf{31.20 \pm 0.06}$	40.58 ± 0.03

Таблица 5.6: Основна задача: подреждане на най-добрите $K = 5$ отговора, върнати от моделите Изв. Инф. и Seq2seq.

предварително обучени word2vec ембединги. Те не са оптимизирани по време на обучението, което помага да се избегнат нежелани влияния върху резултатите (*biases*), както е предложено в Liu et al. (2016); Lowe et al. (2017).

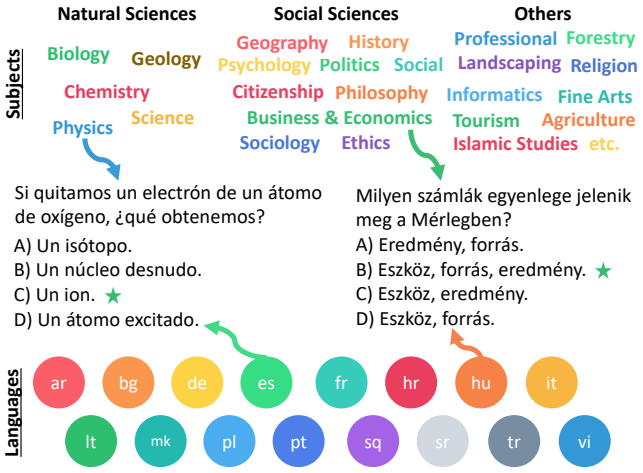
Основна задача: Подреждане на отговорите от множество източници

Следващата стъпка е да се комбинират най-добрите K отговори от различни модели: Изв. Инф. и Seq2seq. Transformer моделът не е включен, тъй като неговият изход е генеративен и съответно подобен на този от Seq2seq модела; освен това, както се вижда от Таблица 5.5 по-горе, той се представя по-лошо от Seq2seq върху избрания корпус. При базовия модел (*Случаен най-добър отговор*) се избира случаен отговор от първите K ($K = 2$) най-добри отговори предложени от моделите . За останалите експерименти с преподреждане се използва $K = 5$. Тези стойности са избрани с кръстосано валидиране (*cross validation*) върху корпуса за обучение, тествайки стойности от 1–5.

Резултатите са показани в Таблица 5.6, където различните представяния са разделени с хоризонтална линия. Първият ред на всяка група съдържа името на модела. След това на четните редове (втори, четвърти и т.н.) са показани резултатите от алчна (*greedy*) стратегия за избор на отговор, докато на нечетните редове са резултатите от стратегия чрез изследване (*exploration strategy*) (семплиране след софтмакс). Тъй като семплирането след софтмакс и произволният избор са стохастични по природа, е включен доверителен интервал (*confidence interval*) от 95% за тях.

5.4 Многоезичност и междуезичност

Тук е представен $E\chi\alpha\mu s$, нов корпус и бенчмарк (*benchmark*) за многоезична и междуезикова оценка на модели и методи за отговаряне на разнообразни научни



Фигура 5.2: Свойства и примери на *Εχαμς*.

въпроси на училищно ниво (вижте Фигура 5.2).

5.4.1 Корпусът *Εχαμς*

Статистики

Εχαμς е събран от официални държавни изпити (матури), подготвени от министерствата на образованието на различни страни. Тези изпити се полагат от ученици, завършващи средно образование, и често изискват знания, научени през целия курс на обучение. Въпросите обхващат много и разнообразни теми и материали, базирани на образователната система на страната. Освен това корпусът не включва само данни от основни учебни предмети като биология, химия, география, история и физика, а и силно специализирани такива като селско стопанство, геология, информатика, както и някои приложни и профилирани предмети. Тези характеристики правят въпросите в корпуса изключително разнообразни и съответно трудно разрешими поради необходимостта от специализирани знания.

Многоезичност Корпусът включва общо 24,143 въпроси на 16 езика от осем езикови семейства. Всеки въпрос има от 3 до 5 възможни отговора (средно 3,96), като само един от тях е правилен. Таблица 5.7 показва разбивка за всеки език, където броят на темите, въпросите и размерът на речника са показани като абсолютни числа, докато дължината на въпроса, дължината на отговора и броят на отговорите са осреднени. Всички статистики за въпросите и вариантите за отговор се измерват в думи.

Паралелни въпроси Някои държави позволяват на учениците да полагат официални изпити на няколко езика. Такива паралелни изпити съществуват и в новосъбрания корпус. По-конкретно, в него има 9 857 паралелни двойки въпроси, разпределени на седем езика, както е показано в Таблица 5.8. Паралелните двойки идват от Хърватия (хърватски, сръбски, италиански, унгарски), Унгария (унгарски, немски, френски, испански, хърватски, сръбски, италиански) и Северна Македония (македонски, албански, турски).

Език	Семейство	#Предмети	Дъл. Въпрос	Дъл. Отговор	#Отговори	#Въпроси	Речник
Албански	Албански	8	15.0	5.0	4.0	1,505	11,572
Арабски	Семитски	5	10.3	3.4	4.0	562	5,189
Български	Балто-славянски	6	13.0	3.3	4.0	2,937	15,127
Хърватски	Балто-славянски	14	14.7	4.1	3.9	2,879	20,689
Френски	Романски	3	18.4	10.5	3.5	318	2,576
Немски	Германски	5	18.3	9.1	3.5	577	4,664
Унгарски	Угро-фински	10	11.6	5.9	3.9	2,267	15,045
Италиански	Романски	12	20.0	5.6	3.9	1,256	9,050
Литовски	Балто-славянски	2	9.7	4.7	4.0	593	5,394
Македонски	Балто-славянски	8	13.4	4.5	4.0	2,075	13,114
Полски	Балто-славянски	1	13.7	4.3	4.0	1,971	18,990
Португалски	Романски	4	19.9	8.6	4.0	924	6,811
Сръбски	Балто-славянски	14	15.4	4.3	3.9	1,637	15,509
Испански	Романски	2	23.0	10.2	3.2	235	2,130
Турски	Тюркски	8	19.5	4.6	4.4	1,964	22,069
Виетнамски	Австроазиатски	6	37.0	6.4	4.0	2,443	6,076
#Езика 16	#Семейства 8	24	17.19	5.08	3.96	24,143	158,942

Таблица 5.7: Статистика за $E\chi\alpha\mu s$. Средната дължина (*дъл.*) на въпроса и отговорите се измерват в брой токени, а размерът на речника се измерва в брой думи.

	de	es	fr	hr	hu	it	mk	sq	sr
de	-								
es	199	-							
fr	253	120	-						
hr	189	134	109	-					
hu	456	159	274	236	-				
it	30	9	15	1,214	99	-			
mk	0	0	0	0	0	0	-		
sq	0	0	0	0	0	0	1,403	-	
sr	40	25	20	1,564	104	1,002	0	0	-
tr	0	0	0	0	0	0	1,222	981	0

Таблица 5.8: Паралелни въпроси между различни езикови двойки.

Множества

Многоезично Целта на тази постановка е да се обучи и оценени даден модел с множество езици. Поради тази причина са нужни нови многоезични *обучаващо*, *вариационно* и *тестово* множество. За да се гарантира, че са включени възможно най-много от езиците, първо въпросите се разделят независимо един от друг за всеки език (L) на подмножества Train_L, Dev_L, Test_L в пропорция 37,5%, 12,5%, и съответно 50%.[2] След това всички подмножества от всеки език са обединени в многоезични – Train_{Mul}, Dev_{Mul}, Test_{Mul}, които по-късно се използват за обучение, валидация и тестване.

Понеже има паралелни данни на няколко езика (обсъдени в Раздел 5.4.1), при разделението паралелните въпроси са разпределени в едно и също множество, така че да няма изтичане на отговори чрез друг език по време на обучението. Броят примери за език и общият брой многоезични множества са показани в първите три колони на Таблица 5.9.[3]

Междуезично Целта на тази постановка е да се изследва способността на модела да прехвърля знанията си от един изходен език L_{src} към нов невиждан

[2]За езици с по-малко от 900 примера има само Test_L.

[3]Понякога групирането на паралелни въпроси в едно и също множество леко нарушава съотношенията зададени за разделяне на данните.

	Многоезично			Междуезично	
Език	**Обучаващо**	**Валид.**	**Тестово**	**Обучаващо**	**Валид.**
Албански	565	185	755	1,194	311
Арабски	-	-	562	-	-
Български	1,100	365	1,472	2,344	593
Хърватски	1,003	335	1,541	2,341	538
Френски	-	-	318	-	-
Немски	-	-	577	-	-
Унгарски	707	263	1,297	1,731	536
Италиански	464	156	636	1,010	246
Литовски	-	-	593	-	-
Македонски	778	265	1,032	1,665	410
Полски	739	246	986	1,577	394
Португалски	346	115	463	740	184
Сръбски	596	197	844	1,323	314
Испански	-	-	235	-	-
Турски	747	240	977	1,571	393
Виетнамски	916	305	1,222	1,955	488
Комбинирано	7,961	2,672	13,510	-	-

Таблица 5.9: Брой примери във всяко множество въз основа на експерименталната постановка.

целеви език L_{tgt}. За да се гарантира по-голям корпус за обучение, моделът е обучен на 80% от L_{src}, и валидиран върху 20% от същия език, а тестовото множество е целият L_{tgt}.[4] Последните три колони на Таблица 5.9 показват броя на примерите, използвани за обучение и валидация на съответния език.

5.4.2 Базови модели

Без допълнително обучение

Извличане на информация Изв. Инф. Този базов модел, използващ Изв. Инф., е взаимстван от Clark et al. (2016) и класира възможните отговори o за всеки въпрос q въз основа на оценката им за релевантност (*relevance*), изчислена от търсачка (*search engine*).[5] По-точно, за всеки възможен отговор o_i се формира заявка, като се конкатенира неговият текст към въпроса ($q + o_i$). Полученият низ се изпраща към търсачката.

Предварително обучен модел като база от знания (БЗ) Тук целта е да се оценят знанията, съдържащи се в модела, като се използва стандартния механизъм за маскиране, приложен в предварителното обучение. Токенизира се всяка двойка въпрос-отговор в поддуми (*subwords*) и след това се заменят всички части от отговора със специалния токен [MASK]. Следвайки нотацията от Devlin et al. (2019), входната последователност може да бъде записана по следния начин:

[CLS] $[Q_1] \ldots [Q_N]$ $[M\_O_1] \ldots [M\_O_M]$ [SEP],

където Q е въпросът, а M_O е маскираният отговор. Следвайки нотацията по-горе се получава оценка за всеки отговор на въпроса въз основа на нормализираната логаритмична вероятност за цялата маскирана последователност (вижте

[4] За да се гарантира, че междуезиковата оценка е сравнима с многоезичната, се използва същото тестово множество език L_{tgt}, Test_{Mul}

[5] Създава се и се използва отделен обърнат индекс за всеки език в ElasticSearch.

	ARC		R12									*Eχαμς*								
Език/Множество	**E**	**C**	**en**	**ar**	**bg**	**de**	**es**	**fr**	**hr**	**hu**	**it**	**lt**	**mk**	**pl**	**pt**	**sq**	**sr**	**tr**	**vi**	**Всички**
Случаен избор	25.0	25.0	25.0	25.0	25.0	29.4	32.0	29.4	26.7	27.7	26.0	25.0	25.0	25.0	25.0	25.0	26.2	23.1	25.0	25.9
IR (Уикипедия)	-	-	-	31.0	29.6	29.3	27.2	32.1	31.9	29.7	27.6	29.8	32.2	29.2	27.5	25.3	31.8	28.5	27.5	29.5
XLM-R on RACE	61.6	45.9	57.4	39.1	43.9	37.2	40.0	37.4	38.8	39.9	36.9	40.5	45.9	33.9	37.4	42.3	35.6	37.1	35.9	39.1
w/ SciENs	**73.6**	51.2	68.4	39.1	44.2	35.5	37.9	37.1	38.5	37.9	39.5	**41.3**	49.8	36.1	**39.3**	42.5	37.4	37.4	35.9	39.6
then on *Eχαμς* (Full)	72.8	**52.6**	**68.8**	**40.7**	**47.2**	**39.7**	**42.1**	**39.6**	**41.6**	**40.2**	**40.6**	40.6	**53.1**	**38.3**	38.9	**44.6**	**39.6**	**40.3**	**37.5**	**42.0**
XLM-R_{Base} (Full)	54.2	36.4	54.6	34.5	35.7	*36.7*	38.3	*36.5*	*35.6*	*33.3*	33.3	33.2	41.4	*30.8*	29.8	*33.5*	32.3	30.4	*32.1*	34.1
mBERT (Full)	*63.8*	*38.9*	*57.0*	34.5	*39.5*	35.3	*40.9*	34.9	35.3	32.7	*36.0*	*34.4*	*42.1*	30.0	29.8	30.9	*34.3*	*31.8*	31.7	*34.6*
mBERT (*Eχαμς* only)	39.6	28.5	35.1	31.9	34.1	30.4	37.9	33.3	32.6	29.3	31.1	31.9	42.4	29.0	28.3	29.9	30.8	25.4	30.0	31.7
XLM-R as KB	30.8	26.2	27.2	31.0	27.2	31.7	37.9	29.9	27.6	29.3	28.0	28.3	23.5	24.6	27.0	25.6	25.4	24.4	24.9	27.0
XLM-R (Full) w/o ctx	45.4	39.2	47.6	30.2	34.8	34.3	30.2	33.0	33.6	33.4	28.5	30.9	37.5	30.0	32.4	36.7	32.1	31.7	30.4	32.8

Таблица 5.10: Обща оценка за всеки език. Първите три колони показват резултатите на ARC Easy (E), ARC Challenge (C) и Regents 12 LivEnv (en). Следващите колони показват точността за всеки език и общите резултати (последната колона Всички) за всички езици. *Всички* е резултатът, осреден за всички *Eχαμς* въпроси.

Формула 5.1).

$$score(O_i) = \frac{1}{|O_i|} \sum_{t \in O_i} \log P_{MLM}(t|Q) \tag{5.1}$$

Модели с доплънително обучение

Целта е да се оцени способността на моделите да предават научни знания между различни езици, когато са обучени допълнително върху такива данни.

5.4.3 Експерименти и резултати

Многоезична оценка

Следващите две групи показват: (*i*) как помага допълнителното обучение на XLM-R върху машинното четене с разбиране на научни въпроси с множествен избор и (*ii*) как се сравняват резултатите от различните модели (XLM-R, XLM-R_{Base} и mBERT). За обучението на тези модели е следвана стандартната схема за тази задача. Първо моделът се обучава върху RACE (Lai et al., 2017) (∼85k английски (EN) въпроси върху документи). След това върху корпусите с научни въпроси от AI2 English (наричани SciENs за по-кратко), включително ∼9k английски въпроси с предоставен подходящ контекст,[6]. Накрая, върху новото, многоезично множество за обучение (вижте Раздел 5.4.1) с извлечени подходящи контексти от Уикипедия. Тази постановка е наречена *Full*. Можем също да видим, че обучението върху корпуса SciENs, който съдържа предимно въпроси от тестове за начални класове от естествените науки, дава само +0,5% подобрение върху *Eχαμς*. Въпреки това виждаме подобрение от +2,4% при многоезично допълнително обучение върху *Eχαμς* (+0,5% за английските корпуси). В третата група се сравняват резултатите от mBERT, XLM-R_{Base} и XLM-R след допълнително обучение. Увеличаването на капацитета на модела води до подобрения: XLM-R отбелязва 7,4% по-висок резултат върху *Eχαμς* и повече от 14% върху английските корпуси в сравнение с базовата му версия (XLM-R_{Base}). Въпреки това mBERT и XLM-R_{Base} имат близка точност, като mBERT има малко предимство в многоезичната постановка. Накрая е представен експеримент с mBERT, обучен само на *Eχαμς*. Както се очаква, точността пада с 3% абсолютна стойност в сравнение с постановката *Full*.

[6]Използват се данните, описани на http://leaderboard.allenai.org/arc/submission/blcotvl7rrltlue6bsv0

Език	A_E	A_{Ch}	R12	de	es	fr	it	pt	bg	hr	lt	mk	pl	sr	hu	sq	tr	vi	ar
en_{all}	73.6*	51.2*	68.4*	35.5*	37.9	37.1	39.5	39.3	44.2	38.5	41.3	49.8	36.1	37.4	37.9	42.5	37.4	35.9	39.1
w/ it	+1.4	+1.3	+1.4	+6.2	**+4.2***	+0.3*	-	-3.7*	+1.2	+4.1	+0.9	+0.8	+1.5	+3.1	+2.8	+0.9	-1.3	**+1.8**	+1.8
w/ pt	+0.1	+1.2	-0.8	+2.2	+2.5*	-2.5*	+1.4*	-	+0.3	0.0	+2.0	+0.8	-0.1	-0.6	-0.6	-1.3	**+1.3**	+0.6	+1.1
w/ bg	+0.6	+0.4	-0.4	**+3.6**	+0.8	+1.6	**+3.4**	-1.9	-	+1.5*	**+2.9***	**+1.6***	+0.1*	**+1.5***	+2.0	**+2.3**	-0.9	-0.8	+0.8
w/ hr	+1.1	**+1.7**	-0.2	+4.8	+3.8	+0.3	+5.8	-2.8	+1.7*	-	+0.2*	-0.1*	+1.2*	+6.7*	+2.8	+1.7	+1.2	+0.5	-0.1
w/ mk	+1.5	-0.5	**+2.2**	+1.0	**+4.2**	-0.3	+2.0	-2.6	+1.8*	**+3.9***	+1.5*	-	+1.9*	0.0*	+2.0	+6.9	+4.8	+0.5	**+4.5**
w/ pl	-2.0	-1.5	-3.1	0.0	+0.4	-2.5	+0.1	-1.3	+1.1*	+1.0*	-0.5*	-0.2*	-	0.0*	-0.4	+0.3	+0.2	-1.4	+0.9
w/ sr	**+1.8**	-0.1	-1.2	+2.6	+5.1	+1.9	+2.8	-0.6	**+2.2***	+6.2*	+0.2*	+1.3*	+1.3*	-	+1.4	-0.4	-0.7	-1.0	+3.2
w/ hu	-0.8	-0.8	-1.0	+7.8	+10.2	+2.8	+1.1	-1.9	+0.7	+0.8	-3.2	+0.1	+0.9	+0.9	-	-0.2	-0.2	-0.6	-1.4
w/ sq	-0.1	+0.3	-1.5	+3.5	-0.5	-0.6	+0.8	+0.9	+0.9	+0.8	+1.0	+3.4	+0.6	+0.6	+1.9	-	+0.4	+0.3	+0.2
w/ tr	-0.5	+1.1	-1.5	+1.5	+3.0	-1.9	+2.3	-3.0	+1.0	+1.0	-2.7	+1.5	+0.2	+1.2	**+2.4**	+3.7	-	-1.0	+1.8
w/ vi	-0.5	+0.4	-0.8	+2.9	+3.4	**+4.1**	+1.1	**+1.1**	+1.5	+1.7	+0.4	+0.4	**+2.1**	0.0	+1.7	+0.8	+1.1	-	+3.4

Таблица 5.11: Междуезична точност върху *Εχαμs* при трансфер без обучение. Първите три колони показват ефективността върху тестовото множество на научните корпуси от AI2 (английски), последвани от оценка за всеки език. Подчертаните стойности маркират езици, които имат паралелни данни с езика източник, а тези със звезда* са от едно и също езиково семейство.

Оценка на знания

Последните два реда от Таблица 5.10 оценяват знанията в най-добрия модел, а именно XLM-R. С *XLM-R като БЗ* (вижте Раздел 5.4.2) виждаме малко подобрение спрямо случайно избиране на отговор: +5% ARC Easy, 2% на R12 и само +1% на *Εχαμs* и ARC Challenge. В допълнение са оценени, знанията, съдържащи се в модела след пълното (*Full*) допълнително обучение, като е изключен съответният контекст за всеки от въпросите (*ctx*). Този модел се представя по-добре от *XLM-R като БЗ*, но все още постига по-ниски общи резултати, което показва, че съхранените знания в модела не са достатъчни и трябва бъдат използвани допълнителни знания от външен източник.

Междуезична оценка

Таблица 5.11 сравнява резултатите от междуезичното трансфериране без допълнително обучение с базов модел, обучен само на английски en_{all}, и XLM-R обучаван върху SciEN. Езиците са подредени първо по езиково семейство и след това по азбучен ред. Моделът се обучава на език (източник) и се тества на всички други (целеви) езици, използвайки множествата, описани в Раздел 5.4.1. Резултатите показват, че ефектът от допълнителното обучение на един език е предимно положителен. Това е забележимо при езиците с подобни лингвистични характеристики с езика към който се трансферира, напр. балто-славянски: bg-sr, hr-mk, pl-mk, sr-bg.

Също така виждаме ползи, когато езика източник съдържа повече въпроси от широко представени и по-трудни теми. Примери за това са експериментите, показващи положителните ефекти от обучението върху виетнамски и македонски като езици, източници; и двете множества съдържат точно такива предмети: биология, история, химия, физика и география.

5.5 Обобщение

В тази глава беше представено изследване на автоматизирано обслужване на клиенти в Twitter, като бяха използвани два типа модели: (*i*) базиран на извличане на информация (Изв. Инф. с BM25) и (*ii*) базирани на генеративни невронни мрежи (Seq2seq с внимание и Трансформатор). Тези модели бяха оценени без

нуждата от човешка преценка, използвайки мерки, базирани на (*i*) припокриване на думи (BLEU@2 и ROUGE-L) и (*ii*) семантика (Embedding Average, Vector Extrema, Greedy Matching). За целта на направените експерименти, данните бяха разделени на база на времето на тяхното публикуване, симулирайки реален сценарии. Направените експерименти показаха, че генеративните невронни модели превъзхождат тези, базирани на извличане на информацията, но не се справят толкова добре, когато има много малко примери от определена тема в данните за обучение. Независимо от това, въпреки че невронните модели показваха добри резултати и бяха в състояние да генерират граматически правилни и подходящи отговори на зададените им въпроси, е видимо, че само данни от диалози между потребител и оператор, не са достатъчни за изграждане на цялостен чатбот за клиентско обслужване. Това може да бъде отдадено на променливия характер на клиентските проблеми – въпреки че отговорите са били коректни при публикуването им, те губят своята актуалност с времето.

Предложена беше нова методология за подреждане на кандидати за отговори за следваща реплика на разговорни агенти. По-точно бяха използвани техники от областта на машинното четене с разбиране (Chen et al., 2017; Seo et al., 2017; Yu et al., 2018), за да се оцени качеството на двойка въпрос-отговор. Предложената методология се състои от две задачи: (i) една спомагателна, целяща да обучи класификатор за определяне на уместността на двойките с помощта на QANet и семплиране на негативни примери, и (ii) основна задача, която пренарежда кандидатите за отговори, използвайки научения модел. Освен това бяха проведени експерименти с различни размери на модела и два типа ембединги: GloVe (Pennington et al., 2014) и ELMo (Peters et al., 2018). Извършените експерименти показаха подобрения в качеството на отговорите от разговорния агент след пренареждането им по отношение на метрики, базирани на припокриване на думи и семантика. Не на последно място се вижда, че изборът на най-високо класирания отговор не винаги е най-добрият вариант. За това беше въведена вероятностна компонента, която има за цел да разнообрази отговорите на агента. Тя все пак дава предимство на популярните отговори, но същевременно взема под внимание техните оценките получени при класирането им.

Накрая беше представен $E\chi\alpha\mu s$– нов предизвикателен междуезичен и многоезичен бенчмарк за отговаряне на научни въпроси на шестнадесет езика и двадесет и четири предмета от матури. Освен това беше предложена нова грануларна оценка на резултатите, която позволява прецизно сравнение между различни езици и училищни предмети. Бяха проведени различни експерименти и анализи с предварително обучени многоезични модели (XLM-R, mBERT). Беше демонстрирана нуждата от разработка на усъвършенствани техники за разсъждение и трансфер на знания, за да бъдат решени някои от въпросите в $E\chi\alpha\mu s$. Публикуваните данни и код ще спомогнат за работа върху многоезични модели, които могат да разсъждават относно отговорите на предизвикателни въпроси от различни научни области.

Глава 6. Заключение и бъдеща работа

6.1 Приноси на дисертацията

Тази дисертация има следните основни научни приноси:

- **Изследване на нови модели и алгоритми**:
  - Беше предложен нов обогатен, предварително обучен езиков модел за съвместно моделиране на задачите за откриване на намерение и запълване на слотове, а именно *Transformer-NLU*. Освен това беше проектиран слой за обединяване на вниманието (*attention pooling*), с цел да се получи по-информативно представяне на намерението, спрямо това получено от специалния стартов токен. Беше подпомогнато запълването на слотовете с специфични за думата характеристики, и предсказаното разпределение от намеренията. Направените експерименти върху два стандартни корпуса показаха, че Transformer-NLU превъзхожда другите алтернативи по всички стандартни мерки, използвани за оценка на PEE задачи.
  - Беше предложен нов подход за обучение с шумни данни, използвайки самоадаптивно обучение и допълнителни тегла във функцията на грешката. Той води до значително по-добър резултат, спрямо силни алтернативи обучени върху изцяло ръчно анотирани данни (Shaar et al., 2021). Отчетено беше подобрение от четири пункта по отношение на MRR, P@1 и MAP@5. Освен това беше демонстрирана полезността на събраните данни, етикетирани с помощта на дистанционно наблюдение (CrowdChecked). Нещо повече, предложеният метод в комбинация с данните от CrowdChecked, постигна по-добри резултати (с две точки) спрямо най-съвременните алтернативи. Моделът постигна MAP@5 от 90.3 върху корпуса CheckThat '21.
  - Беше проектиран цялостен подход за решаване на задачата за четене с разбиране на въпроси с множествен избор за езици с малко ресурси. Предложеният модел се базира на многоезичен BERT (Devlin et al., 2019), който е обучен върху голям корпус на английски език за четене с разбиране и източници на знания от отворен домейн (Уикипедия). Основните експерименти бяха фокусирани върху оценка на модела при трансфер на знания без допълнително обучение.
  - Разработен беше подход за автоматизиране на обслужването на клиенти в Twitter, използвайки два типа модели: (*i*) базиран на извличане (Изв. Инф. с BM25) и (*ii*) базиран на генеративни невронни мрежи (seq2seq с внимание и Transformer). Те бяха оценени без нуждата от човешка преценка, използвайки мерки, базирани на (*i*) припокриване на думи (BLEU@2 и ROUGE-L) и (*ii*) семантика (Embeeding Average,

Greedy Matching и Vector Extrema). Направените експерименти показаха, че генеративните невронни модели превъзхождат тези, базирани на извличане на информацията, но не се справят толкова добре, когато има много малко примери от определена тема в данните за обучение.

– Разработен беше нов подход за подреждане на кандидати за отговори от разговорни агенти. По-конкретно, бяха използвани техники от областта на машинното четене с разбиране (Chen et al., 2017; Seo et al., 2017; Yu et al., 2018), за да се оцени качеството на двойка въпрос-отговор. Предложената методология се състои от две задачи: (i) една спомагателна, целяща да обучи класификатор за определяне на уместността на двойките с помощта на QANet и семплиране на негативни примери, и (ii) основна задача, която пренарежда кандидатите за отговори, използвайки научения модел. Проведени бяха експерименти с различни размери на модела и два типа ембединги: GloVe (Pennington et al., 2014) и ELMo (Peters et al., 2018). Те показаха подобрения в качеството на отговорите от разговорния агент след пренареждането им. За оценка бяха използвани автоматични метрики, базирани на припокриване на думи и семантика.

– Създаден беше нов, предизвикателен междуезичен и многоезичен бенчмарк за отговаряне на научни въпроси от гимназиални изпити. Бяха оценени способностите на най-съвременните модели за трансфер без обучение и междуезичен трансфер в многоезична постановка. Показано беше, че предварителното обучение на големи английски корпуси извън домейна може да помогне на модела да научи задачата, но допълнителни подобрения могат да бъдат постигнати само чрез многоезични данни от целевия домейн.

– Проведени бяха различни експерименти и анализи върху предварително обучени многоезични модели (XLM-R, mBERT). Беше демонстрирана нуждата от по-добри умения за разсъждение и трансфер на знания, за да се решат някои от въпросите от $E\chi\alpha\mu s$.

- **Създаване на нови корпуси** :

  – Беше събран нов български корпус за четене с разбиране на въпроси с избираем отговор с 2 633 въпроса от матури за дванадесети клас по история и биология и онлайн изпити по история без пояснителен контекст.

  – Беше събран $E\chi\alpha\mu s$– нов предизвикателен междуезичен и многоезичен бенчмарк за отговаряне на научни въпроси от матури на шестнадесет езика и двадесет и четири предмета от гимназиални изпити. Освен това беше предложена нова грануларна оценка, която позволява прецизно сравнение на точността на моделите между различни езици и училищни предмети.

  – Беше създаден CrowdChecked– голям по размер корпус за откриване на вече проверени твърдения, с повече от 330 000 двойки туитове и съответстващите им статии за проверка на факти, публикувани от хора от тълпата, проверяващи факти. Допълнително бяха проучени две техники за етикетиране на двойки туит-статия с помощта на дистанционен надзор, базиран на сходството на Жакарад, и предсказанията от модел тип невронна мрежа. По този начин бяха създадени нови обучаващи множества с размер от 3,5K-50K примера.

6.2 Посоки за бъдещи изследвания

Модулираните (ориентирани към задачи) разговорни агенти осигуряват голяма гъвкавост по отношение на обучението на модела и позволяват лесно добавяне на нови или замяна на съществуващи модули към агента. Тази гъвкавост обаче носи няколко ограничения. Първо, има прекъсване на връзката между различни компоненти (модели) както по време на обучение, така и по време на предсказание, което от своя страна води до натрупване на грешки след всяка последователна стъпка от системата. И второ, включването на твърде много компоненти може да увеличи изчислителните разходи, следователно внедряването на системата за диалог може да стане невъзможно. Тук са очертани няколко обещаващи насоки за бъдещи изследвания:

- В краткосрочен план цялостно диференцируемите архитектури базирани на комбинация от йерархични невронни мрежи, многозадачно обучение и многомоделно разпространение на грешките, могат да бъдат стъпка напред в тази посока.
- В дългосрочен план, архитектурите с единичен модел, базирани на цялостни генеративни архитектури, могат да бъдат добра алтернатива на системите с множество модели, дори в сценарии, ориентирани към задачи.
- Дори след постигнатите скорошни успехи в разработването на модели с увеличен капацитет от милиарди обучаеми параметри, те все още са уязвими както към етични, така и към практически рискове (Bender et al., 2021; Bommasani et al., 2021). Въпреки това е ясно, че има нужда от повече изследвания и по-добри цялостни генеративни модели, за да могат те да бъдат внедрени в динамични сценарии от реалния свят. Някои посоки са:
  - Разработване на ефективни механизми за актуализиране на фактологичните знания, съхранени в самия модел (De Cao et al., 2021),
  - Внедряване на допълнителни знания (*knowledge grounding*) (Zhao et al., 2020),
  - Разработване на процедури за автоматично премахване на предразсъдъци в модела (Guo et al., 2022), за да се гарантира, че чатботовете предлагат подходящи и фактологични отговори.
  - И накрая, трябва бъдат бъдат подобрени механизмите, които предпазват моделите от злонамерени участници (Hancock et al., 2019; Vanderlyn et al., 2021).
- Обяснимостта на моделите се превръща във важна изследователска област в ОЕЕ (Danilevsky et al., 2020). Някои интересни бъдещи насоки са: методи, които се фокусират върху обяснението на веригата на разсъжденията (Yang et al., 2018; Das et al., 2018); формиране на отговори в дълга форма с подробни обяснения въз основа на параграфи с доказателства (Kwiatkowski et al., 2019; Fan et al., 2019) и допълнителното им обогатяване (Schick et al., 2022) с автоматични редакции, добавяне на текст, цитати и т.н. или получаване на обяснения на ниво токен (Li and Yao, 2021; Arora et al., 2022).

Приложения А–В

- **Приложение А** обсъжда хиперпараметрите, използвани за обучение на моделите, предложени в Раздел 4.3 *Извличане на отговор от колекция с*

обяснения. Описани са инструкциите за анотация, демографските данни за анотаторите, съгласието между тях и е направен анализ на разликите в техните анотации.

- **Приложение В** предоставя дефиниции на всички теми, включени в корпуса *Eχαμs* (Раздел 5.5 *Многоезичност и междуезичност*). Описани са процедурата за фина настройка и хиперпараметрите на моделите.

Декларация за оригиналност

Декларирам, че настоящият дисертационен труд съдържа оригинални резултати, получени при проведени от мен научни изследвания, с подкрепата на научния ми ръководител и съавтори. Резултатите, които са получени, описани и/или публикувани от други учени са надлежно и подробно цитирани в библиографията. Настоящата работа не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Подпис:

Библиография

Siddhant Arora, Danish Pruthi, Norman Sadeh, William W. Cohen, Zachary C. Lipton, and Graham Neubig. 2022. Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5277–5285.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations*, ICLR '15, San Diego, California, USA.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 215–236. Springer.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, New York, USA.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 1870–1879, Vancouver, Canada.

Anton Chernyavskiy, Dmitry Ilvovsky, Pavel Kalinin, and Preslav Nakov. 2022. Batch-Softmax Contrastive Loss for Pairwise Sentence Scoring Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 116–126, Seattle, Washington, USA.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2021. Aschern at CLEF CheckThat! 2021: Lambda-Calculus of Fact-Checked Claims. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 484–493.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, AAAI '16, pages 2580–2586, Phoenix, Arizona, USA.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2018. Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning. In *International Conference on Learning Representations*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing Factual Knowledge in Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A Survey of Natural Language Generation. *ACM Comput. Surv.* Just Accepted.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471, Florence, Italy.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the CLEF-2019 CheckThat! Lab: automatic identification and verification of claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 301–321. Springer.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots*. IEEE Xplore.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-Gated Modeling for Joint Slot Filling and Intent Prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757, New Orleans, Louisiana.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland.

Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Vivian Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-Domain Joint Semantic Frame Parsing using Bi-directional RNN-LSTM. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association*, INTERSPEECH '16, pages 715–719, San Francisco, USA.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, CoNLL '19, pages 493–503, Hong Kong, China.

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, AACL-IJCNLP '22, Online.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2018. Towards Automated Customer Support. In *Proceedings of the 18th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '18, pages 48–59, Varna, Bulgaria.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019a. Beyond English-Only Reading Comprehension: Experiments in Zero-shot Multilingual Transfer for Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '19, pages 447–459, Varna, Bulgaria.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019b. Machine Reading Comprehension for Answer Re-Ranking in Customer Support Chatbots. *Information*, 10(3).

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2020a. Enriched Pre-trained Transformers for Joint Slot Filling and Intent Detection. *arXiv preprint arXiv:2004.14848*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020b. EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 5427–5444, Online.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: The First-Ever End-to-End Fact-Checking System. *Proc. VLDB Endow.*, 10(12):1945–1948.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language: Proceedings of a Workshop*, Hidden Valley, Pennsylvania.

Matthew Henderson, Rami Al-Rfou, B. Strope, Yun-Hsuan Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *ArXiv 1705.00652*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Lang Huang, Chao Zhang, and Hongyang Zhang. 2020. Self-Adaptive Training: beyond Empirical Risk Minimization. In *Advances in Neural Information Processing Systems*, volume 33.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 785–794, Copenhagen, Denmark.

Yangming Li and Kaisheng Yao. 2021. Interpretable NLG for Task-oriented Dialogue Systems with Heterogeneous Rendering Machines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13306–13314.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics*, ACL '04, pages 605–612, Barcelona, Spain.

Bing Liu and Ian Lane. 2016. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association*, INTERSPEECH '16, pages 685–689, San Francisco, USA.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 2122–2132, Austin, Texas, USA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 1116–1126, Vancouver, Canada.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '14, pages 55–60, Baltimore, Maryland.

Simona Mihaylova, Iva Borisova, Dzhovani Chemishanov, Preslav Hadzhitsanev, Momchil Hardalov, and Preslav Nakov. 2021. DIPS at CheckThat! 2021: Verified Claim Retrieval. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 558–571.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119.

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie, and Dong Yu. 2019. Improving Question Answering with External Knowledge. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, MRQA '19, pages 27–37, Hong Kong, China.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, NAACL-HLT '18, pages 2227–2237, New Orleans, Louisiana.

Popat, Kashyap and Mukherjee, Subhabrata and Strötgen, Jannik and Weikum, Gerhard. 2016. Credibility Assessment of Textual Claims on the Web. In *CIKM*.

Albert Pritzkau. 2021. NLytics at CheckThat! 2021: Multi-class fake news detection of news articles and domain identification with RoBERTa - a baseline model. In *CLEF (Working Notes)*, pages 572–581.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2078–2087, Hong Kong, China.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1807–1816, Online.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/languageunderstandingpaper.pdf.

Kiran Ramesh, Surya Ravishankaran, Abhishek Joshi, and K. Chandrasekaran. 2017. A Survey of Design Techniques for Conversational Agents. In *Information, Communication and Computing Technology*, pages 336–350, Singapore.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. PEER: A Collaborative Language Model. *arXiv preprint arXiv:2208.11663*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the 2017 International Conference on Learning Representations*, ICLR '17, Toulon, France.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online.

Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, Giovanni Da San Martino, Tamer Elsayed, and Preslav Nakov. 2021. Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates. In *CLEF (Working Notes)*, pages 393–405.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training Very Deep Networks. In *Advances in Neural Information Processing Systems*, volume 28.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Improving Machine Reading Comprehension with General Reading Strategies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 2633–2643, Minneapolis, Minnesota, USA.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-range Reasoning for Machine Comprehension. *arXiv preprint arXiv:1803.09074*.

Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A knowledge graph of fact-checked claims. In *International Semantic Web Conference*, pages 309–324. Springer.

Lindsey Vanderlyn, Gianna Weber, Michael Neumann, Dirk Väth, Sarina Meyer, and Ngoc Thang Vu. 2021. "It seemed like an annoying woman": On the Perception and Ethical Considerations of Affective Language in Text-Based Conversational Agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 44–57, Online.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, NIPS '17, pages 5998–6008, Long Beach, CA, USA.

Nguyen Vo and Kyumin Lee. 2019. Learning from Fact-Checkers: Analysis and Generation of Fact-Checking Language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 335–344.

Nguyen Vo and Kyumin Lee. 2020. Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding. *ACM Comput. Surv.* Just Accepted.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Jason D Williams and Geoffrey Zweig. 2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*.

Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 78–83, Olomouc, Czech Republic. IEEE.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2369–2380, Brussels, Belgium.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In *Proceedings of the 2018 International Conference on Learning Representations*, ICLR '18, Vancouver, Canada.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip Yu. 2019. Joint Slot Filling and Intent Detection via Capsule Neural Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5259–5267, Florence, Italy.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 3377–3390, Online.