FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Bakalářská práce

Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek

Martin Reich

PLZEŇ

2024

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Bakalářská práce

Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek

Martin Reich

Vedoucí práce
Doc. Ing. Roman Mouček, Ph.D.

PLZEŇ 2024

Citace v seznamu literatury:
REICH, Martin. *Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek*. Plzeň, 2024. Bakalářská práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky. Vedoucí práce Doc. Ing. Roman Mouček, Ph.D.

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2024/2025

Studijní program: Informatika a výpočetní technika
Forma studia: Prezenční
Specializace/kombinace: Informatika (INF18bp)

Podklad pro zadání BAKALÁŘSKÉ práce studenta

Jméno a příjmení:	**Martin REICH**
Osobní číslo:	**A22B0123P**
Adresa:	Plovární 1458/21, Plzeň – Jižní Předměstí, 30100 Plzeň 1, Česká republika
Téma práce:	Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek
Téma práce anglicky:	An integrated system for automating the recording, editing and publication of lectures
Jazyk práce:	Čeština
Související osoby:	Doc. Ing. Roman Mouček, Ph.D. (Vedoucí) Katedra informatiky a výpočetní techniky

Zásady pro vypracování:

1. Seznamte se současnými metodami a možnostmi převodu řeči do textu a zpět a metodami parafrázování
2. Seznamte se současnými systémy, nástroji a knihovnami pro transkripci řeči na text, převod textu na řeč a parafrázování
3. Navrhněte integrovaný systém, který dokáže automatizovat převod zvukového a obrazového záznamu přednášky v češtině či angličtině do textu, tento následně parafrázovat a zpětně převést do zvukové a obrazové podoby přednášky
4. Navržený systém dle bodu 3 implementujte.
5. Ověřte výsledné řešení na dostatečném počtu přednášek v českém i anglickém jazyce.

Seznam doporučené literatury:

Dodá vedoucí práce

Podpis studenta: Datum:

Podpis vedoucího práce: Datum:

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného akademického titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Západočeská univerzita v Plzni má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Plzni dne 31. prosince 2024

. . . . . .

Martin Reich

Abstrakt

TODO

Abstract

TODO in english

Klíčová slova

Převod řeči na text • AI • Převod textu na řeč • Parafrázování • Automatizace

Poděkování

Rád bych tímto poděkoval vedoucímu bakalářské práce Doc. Ing. Romanu Moučkovi, Ph.D. za pomoc a odborné vedení při vypracování této práce.

Obsah

Analýza problému 1

1.1 Studium současných metod převodu řeči do textu a zpět, včetně parafrázování

V první fázi je nutné porozumět současným technologiím pro:

- Převod řeči na text: Mezi nejznámější metody patří systémy založené na hlubokých neuronových sítích, konkrétně na architekturách jako jsou konvoluční neuronové sítě (CNN) a rekurentní neuronové sítě (RNN), nebo i pokročilejší transformery. Pro hodnocení efektivity a vhodnosti těchto metod se používají různé metriky. K těm nejčastěji využívaným patří standardní metriky hlubokého učení, jako je přesnost (accuracy), F1-skóre, recall (senzitivita nebo míra pravdivých pozitiv, TPR), precision (pozitivní prediktivní hodnota) a specificity (míra pravdivých negativ, TNR).

- Převod textu na řeč: Nejnovější metody TTS využívají neuronové sítě, jako třeba TTS model od OpenAI, ElevenLabs modely, Google TTS model atd.. Tyto modely zajišťují přirozenější syntézu řeči a mohou pracovat s intonací a rytmem, což je klíčové pro parafrázované texty. K hodnocení výkonnosti těchto modelů se využívají objektivní a subjektivní metriky jako např.: Mel cepstral distortion (MCD), měřící podobnost syntetizovaného a přirozeného zvuku. Root mean square error (RMSE) pro logaritmus základní frekvence (f0). Voiced/unvoiced (V/UV) error rate pro analýzu správnosti rozhodnutí o znělosti. Gross pitch error (GPE) pro hodnocení odchylek výšky hlasu. Mean opinion score (MOS), kde posluchači hodnotí kvalitu syntézy na stupnici od 0 do 5. Více zde [KS23].

- Parafrázování: Parafrázování textu je možné provádět pomocí modelů NLP(natural language processing, zpracování přirozeného jazyka), které dokážou přepsat text jinými slovy, aniž by změnily význam. Zde se využívají transformer modely, například T5 nebo GPT-4. Při parafrázování je výzvou zachovat sro-

zumitelnost a přirozenost textu, zejména pokud se bude převádět zpět na řeč.

1.2 Průzkum dostupných modelů, nástrojů a knihoven a jejich vyzkoušení

Druhá fáze zahrnuje detailní průzkum stávajících nástrojů, modelů a knihoven, které podporují každý krok procesu a jejich vyzkoušení. Pro STT:

- Modely a nástroje:
  - OpenAI Whisper: Open-source model s vysokou přesností a podporou více jazyků.
  - Google Speech-to-Text API: Komerční služba s možností přizpůsobení pro konkrétní domény.
  - DeepSpeech: Open-source model od Mozilly vhodný pro vlastní nasazení.
  - Microsoft Azure Speech-to-Text: Cloudová služba s podporou integrace do enterprise aplikací.
- Knihovny:
  - SpeechRecognition: Python knihovna pro jednoduchou práci s STT.
  - PyTorch a TensorFlow: Frameworky využívané pro implementaci a trénink vlastních modelů STT.

Pro TTS:

- Modely a nástroje:
  - ElevenLabs Speech Synthesis: Pokročilý model s podporou práce s emocemi a intonací.
  - Google TTS API: Komerční služba s realistickou syntézou řeči a možnostmi přizpůsobení.
  - Microsoft Azure Text-to-Speech: Služba s vysokou kvalitou syntézy a jazykovou podporou.
- Knihovny:
  - PyTorch a TensorFlow: Frameworky využívané pro implementaci a trénink vlastních modelů TTS.

Pro parafrázování:

- Modely a nástroje:
  - OpenAI GPT-4: Výkonný generativní model schopný sofistikovaného parafrázování.
  - BART (Facebook): Transformer zaměřený na generování textu a parafrázování.
  - T5 (Text-to-Text Transfer Transformer): Univerzální model od Googlu pro různé NLP úkoly, včetně parafrázování.
- Knihovny:
  - Hugging Face Transformers: Knihovna podporující širokou škálu NLP modelů, včetně BART, T5 a GPT.
  - PyTorch a TensorFlow: Frameworky využívané pro implementaci a trénink vlastních modelů.

TODO Doptat se jak udělat úspěšnost modelů

1.3 Návrh integrovaného systému a jeho implementace

Dále je potřeba si rozmyslet a navrhnout systém, který:

- Automatizuje převod mluveného slova na text: Tento krok zahrnuje přepis audia nebo video záznamů přednášky na text.
- Parafrázuje přepsaný text: Po přepisu je text předán k parafrázování, aby výsledný text byl přeformulovaný.
- Převádí parafrázovaný text zpět na řeč: Po parafrázování je třeba převést text na řeč, která by měla být pokud možno přirozená a příjemná k poslechu.
- Synchronizuje obraz a zvuk: Nakonec je potřeba synchronizovat nový zvuk s obrazem, který může být originální nebo nový ve formě animace či avatara.

Po navrhnutí takovéhoto systému je potřeba ho implementovat(viz. implementace**Dodat odkaz po přidání implementace**)

1.4 Testování systému

Na závěr je potřeba zvolit vhodnou metodu pro testování navrženého a implementované systému. Tento krok také zahrnuje testování podle kritérií na reálných přednáškách v českém a anglickém jazyce. Mezi tyto kritéria může patřit:

- Kvalita: Testování přesnosti rozpoznávání řeči, přirozenosti parafrázování a srozumitelnosti výstupu.
- Uživatelské hodnocení: Testování s cílovými uživateli, kteří posoudí užitečnost a přirozenost celého procesu.
- Výkon: Hodnocení rychlosti a efektivity systému, aby byla zajištěna schopnost zpracovat delší přednášky bez výrazného zpoždění.

Převod řeči do textu 2

Převedení řeči do textu (Speech-to-Text, STT) je oblast, která se zabývá automatickým rozpoznáváním mluveného slova a jeho převodem na psaný text. Je to multidisciplinární obor zahrnující akustiku, jazykovědu, statistiku a strojové učení. Současné technologie pro převod řeči na text jsou poháněny pokročilými modely strojového učení. Tyto modely využívají akustické modely (AM) a jazykové modely (LM), které spolupracují na přesné interpretaci mluvené řeči. Hlavními pilíři moderních přístupů jsou hluboké neuronové sítě, transformery, a techniky přenosového učení (DTL), které umožňují přizpůsobení modelů specifickým podmínkám. **TODO Doptat se tam nechat popis jak to funguje(ale udělat to profesionálnější)**

2.1 Metody

Současné metody, které se používají k převodu reči na text. **TODO Doptat se jestli mám popsat modely obecně nebo v kontextu s STT**

2.1.1 Umělé neuronové sítě

Umělé neuronové sítě hrají klíčovou roli v systémech pro převod řeči na text. Tyto sítě jsou inspirovány strukturou a funkcí lidského mozku a napodobují procesy rozhodování prostřednictvím propojení umělých neuronů. Každá neuronová síť se skládá z vrstev uzlů zahrnujících vstupní, skryté a výstupní vrstvy. V SST neuronové sítě dokáží analyzovat složité akustické vzory, které vznikají z mluvené řeči, a převádět je do strukturovaného textu. Tyto modely se učí z velkých množství tréninkových dat, což jim umožňuje zlepšovat přesnost a přizpůsobit se různým jazykovým kontextům a akustickým prostředím. Optimalizované neuronové sítě mohou provádět úlohy, jako je rozpoznávání řeči, rychle a efektivně, což umožňuje nasazení v aplikacích, jako jsou virtuální asistenti nebo titulkovací systémy.
Výhody:

- Přesnost: Díky schopnosti identifikovat složité akustické a jazykové vzory jsou neuronové sítě schopny dosahovat vysoké přesnosti rozpoznávání.

- Adaptabilita: Sítě lze trénovat pro různé jazyky, akcenty a aplikace, což je činí univerzálními.
- Všestrannost: Neuronové sítě podporují různé přístupy, jako je klasická segmentace na fonémy i moderní end-to-end modely.

Nevýhody:

- Požadavky na tréninková data: Neuronové sítě potřebují velké množství kvalitních tréninkových dat, aby byly schopny efektivně se učit a generalizovat na nová data.
- Trénink neuronových sítí může být časově náročný a vyžaduje značné výpočetní zdroje, zejména při práci s velkými datovými sadami nebo složitými architekturami.

2.1.2 Hluboké učení

Hluboké učení je podmnožina strojového učení, která využívá více vrstvé neuronové sítě, nazývané hluboké neuronové sítě, k simulaci složitého rozhodovacího procesu lidského mozku. Hlavní rozdíl mezi hlubokým učením a strojovým učením spočívá ve struktuře architektury základní neuronové sítě. Tradiční modely strojového učení, které nejsou hluboké, používají jednoduché neuronové sítě s jednou nebo dvěma výpočetními vrstvami. Naopak modely hlubokého učení mají tři a více vrstev, často stovky nebo dokonce tisíce vrstev.

2.1.3 Rekurentní neuronové sítě

Rekurentní neuronové sítě (RNN) jsou modely hlubokého učení a jsou schopné zpracovávat sekvence dat tím, že mají vnitřní paměť, která uchovává informace o předchozích vstupech. Tato vlastnost jim umožňuje modelovat časové závislosti a vztahy v datech. Na rozdíl od tradičních neuronových sítí, které zpracovávají vstupy nezávisle, RNN zohledňují kontext a historii dat. Výhody:

- Zpracování sekvenčních dat: RNN jsou ideální pro úkoly, kde je pořadí dat důležité.
- Flexibilita: Mohou zpracovávat vstupy a výstupy různé délky (např. různé délky vět).

Nevýhody:

- Vanishing gradient: Při trénování mohou gradienty zmizet, což brání učení dlouhodobých závislostí.

- Vysoké výpočetní nároky: RNN mohou být náročné na výpočetní výkon, zejména při práci s dlouhými sekvencemi.

Dodat obrázek rekuretní neuronové sítě

2.1.4 Transformer

Transformery jsou modely hlubokého učení a používají mechanismus vlastní pozornosti, který umožňuje modelu vážit různá slova v sekvenci na základě jejich relevance. To znamená, že model může posoudit, která slova mají vliv na ostatní slova, což je klíčové pro porozumění kontextu. Jelikož transformery nemají vnitřní strukturu pro zpracování sekvencí (na rozdíl od RNN 2.1.3), používají se k nim pozicové kódování, aby modely mohly rozpoznat pořadí slov v sekvenci. Tato kódování přidávají k vektorům slov informace o jejich pozici v sekvenci. Výhody:

- Paralelizace: Transformery umožňují paralelní zpracování vstupních dat, což urychluje trénink ve srovnání s RNN.
- Zachycení dlouhodobých závislostí: Díky mechanismu pozornosti jsou schopny efektivně sledovat vztahy mezi slovy na delší vzdálenosti.
- Flexibilita: Lze je snadno aplikovat na různé úkoly a adaptovat je pro konkrétní aplikace.

Nevýhody:

- Vysoké nároky na paměť: Vzhledem k tomu, že transformery zpracovávají celou sekvenci najednou, mohou mít vysoké požadavky na paměť, zejména při práci s dlouhými sekvencemi.
- Potřeba velkých dat: K dosažení dobrého výkonu vyžadují transformery velké množství tréninkových dat.

2.1.5 Konvoluční neuronové sítě

Konvoluční neuronové sítě (CNN) jsou specifickým typem neuronových sítí, které se ukázaly jako velmi efektivní při analýze vizuálních dat, jako jsou obrázky a videa. Tyto sítě se široce používají v oblastech, jako je počítačové vidění, rozpoznávání obrazů, analýza videí a dokonce i v zpracování přirozeného jazyka. CNN má tři hlavní typy vrstev.

- Konvoluční vrstva: Tato vrstva aplikuje konvoluční operaci na vstupní data pomocí malých filtrů(filtr pokrývá jen část vstupniách dat). Hodnoty(váhy) ve filtru se vynásobí s hodnotami na vstupních datech, preš který je filtr "umístěn".

Filtr se poté přesune, aby vytvořil reprezentaci nové části vstupních dat, a proces se opakuje, dokud nejsou pokrytá celá vstupní data.

- Pooling vrstva: Podobně jako u konvoluční vsrtvy, tato vrstva aplikuje filtry na vstupní data. Jediný rozdíl je, že filtr nemá žádné váhy a ze vstupních dat se vezme hodnotu se největší nebo průměrnou hodnotou.
- Plně propojená vrstva: Na konci CNN se obvykle nachází jedna nebo více plně propojených vrstev, které kombinují výstupy z předchozích vrstev a předpovídají finální výstup.

Výhody:

- Efektivita: CNN dokáží zpracovávat velké množství dat s menšími výpočetními nároky díky sdílení vah a redukci rozměrů pomocí pooling vrstev.
- Vysoká přesnost: Tyto sítě dosahují vynikajících výsledků v úlohách klasifikace obrazů a detekce objektů.

Nevýhody:

- Potřeba velkého množství dat: Pro úspěšné trénování vyžadují CNN velké množství označených tréninkových dat.
- Omezené schopnosti na neobvyklých datech: Modely CNN mohou mít potíže s generalizací na atypická data, která se výrazně liší od tréninkových dat.

2.2 Evaluační metriky

TODO zeptat se jestli to mám přidat, nebo nechat.

2.3 Vybrané modely pro řešení daného problému

TODO dopsat vybrané modely jejich výhody a nevýhody

2.3.1 Testování a výsledky modelů

TODO dodat graf/y s výsledky testování modelů podle rychlosti a přesnosti a napsat který model bude použit k řešení probmlému.

Parafrázování 3

Parafrázování je proces přeformulování textu nebo myšlenek, při kterém zachováme původní význam, ale změníme slovní formulaci a strukturu. Cílem parafrázování je převést informace jinými slovy, často pro lepší pochopení, přizpůsobení určitému publiku nebo aby se text stal originálnějším a nevznikl tak problém s plagiátorstvím.

3.1 Tradiční metody parafrázování

Tradiční přístupy k parafrázování jsou metody, které nevyužívají neuronové sítě, a spoléhají na pravidla, tezaury[1] nebo statistické modely. Tyto přístupy byly v počátcích výzkumu parafrázování široce využívány a poskytly základ pro pozdější vývoj pokročilejších technik.

3.1.1 Pravidlové přístupy

Tyto metody parafrázování využívají pravidla, která mohou být buď ručně vytvořena, nebo automaticky generována. V počátečních výzkumech byla pravidla většinou sestavována manuálně, což bylo časově náročné. Následně byly navrženy postupy umožňující automatizované získávání těchto pravidel.

3.1.2 Přístupy založené na tezaurech

Tento přístup generuje parafráze nahrazením některých slov ve zdrojových větách jejich synonymy získanými z tezauru. Proces zahrnuje nejprve extrakci všech synonym pro nahrazovaná slova a následný výběr optimálního kandidáta podle kontextu zdrojové věty.

[1] Řízený slovník termínů (klíčových slov), které se používají pro vyjádření obsahu dokumentu.

3.1.3 Přístupy založené na statistickém strojovém překladu (SMT)

Tento přístup vychází z myšlenky, že parafrázování lze považovat za speciální případ strojového překladu, konkrétně za monolingvální překlad. Model statistického překladu hledá nejlepší překlad zdrojového textu na základě pravděpodobnostního a jazykového modelu. Analogicky při parafrázování hledá model nejlepší parafrázi zdrojového textu za pomocí pravděpodobnostního a jazykového modelu.

3.2 Neuronové přístupy

3.3 Evaluační metriky

Pro hodnocení generování parafrází se běžně používají dva základní typy evaluačních metrik: automatické hodnocení a hodnocení lidmi.
Automatické hodnocení – Mezi nejběžněji používané metriky pro automatické hodnocení generování parafrází patří:

- BLEU (Papineni et al., 2002), původně vyvinutý pro hodnocení strojového překladu.
- METEOR (Denkowski a Lavie, 2014), který řeší slabiny BLEU v měření sémantických ekvivalentů a lépe koreluje s hodnocením lidí na úrovni vět a segmentů.
- ROUGE (Lin, 2004), metrika založená na recallu, původně vytvořená pro hodnocení sumarizace textu.
- TER (Snover), která měří počet úprav potřebných k tomu, aby lidský překladatel změnil překlad tak, aby přesně odpovídal referenčnímu překladu. Výsledky TER jsou v rozsahu 0-1 a obvykle se prezentují jako procenta, přičemž nižší hodnoty jsou lepší.

Hodnocení lidmi – Automatické metriky se zaměřují hlavně na překryvy n-gramů, což nezohledňuje význam. Proto je pro přesnější a kvalitativní hodnocení generovaného výstupu využíváno hodnocení lidmi. Lidé hodnotí parafráze podle různých kvalitativních dimenzí, jako je podobnost, srozumitelnost a plynulost. I když je hodnocení lidmi časově náročné a nákladné, poskytuje lepší a reprezentativnější obraz o kvalitě generovaného výstupu.

3.4 Vybrané modely pro řešení daného problému

TODO dopsat vybrané modely jejich výhody a nevýhody

3.4.1 Testování a výsledky modelů

TODO dodat graf/y s výsledky testování modelů podle rychlosti a napsat který model bude použit a proč k řešení probmlému.

Zkratky 4

TTS	text-to-speech
RNN	rekuretní neuronová síť
CNN	konvoluční neuronová sít
STT	speech-to-text
TPR	true positive rate
TNR	true negative rate
NLP	natural language processing

Tabulka 4.1: Tabulka zkratek

Bibliografie

[KS23] KAUR, Navdeep; SINGH, Parminder. Conventional and contemporary approaches used in text to speech synthesis: a review. *Artificial Intelligence Review*. 2023, roč. 56. Dostupné také z: https://link.springer.com/article/10.1007/s10462-022-10315-0.

[KHH24] KHEDDAR, Hamza; HEMIS, Mustapha; HIMEUR, Yassine. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*. 2024, roč. 109. Dostupné také z: https://www.sciencedirect.com/science/article/pii/S1566253524002008.