Západočeská univerzita v Plzni
Fakulta filozofická

Diplomová práce

Jazykový model Wittgensteinovy pozdní filosofie

Bc. Jan Tvrz

Západočeská univerzita v Plzni
Fakulta filozofická

Katedra filozofie
Studijní program: Filozofie pro umělou inteligenci

Diplomová práce

Jazykový model Wittgensteinovy pozdní filosofie

Bc. Jan Tvrz

Vedoucí práce:
Mgr. Radek Schuster, Ph.D.
Katedra filozofie
Fakulta filozofická Západočeské univerzity v Plzni

Plzeň 2023

Prohlašuji, že jsem práci zpracoval(a) samostatnì a použil(a) jen uvedených pramenù a literatury.

Plzeň, spren 2023 ……

Obsah:

Úvod

Práce má sloužit jako propojení dvou oborů. Oboru Zpracování přirozeného jazyka a oboru Filosofie. Jedná se tak o kulminaci mého studijního programu Filosofie pro umělou inteligenci. Mým cílem je popsat principy modelování přirozeného jazyka do takové míry, aby se v oboru mohl začít orientovat i laik. Na druhou stranu chci také popsat myšlenky Wittgensteinovy pozdní filosofie, aby byly pochopitelné pro někoho, kdo nemá filosofické školení. Třetím a finálním cílem je popsat postup, který by mohl vést ke tvorbě jazykového modelu filosofa za využití před-trénovaných transformerů.

Tato práce je strukturována do tří sekcí. Sekce A je koncipována jako seznámení s oborem modelování přirozeného jazyka. Jejím cílem je představení stěžejních konceptů a motivací pro tento obor. V první kapitole je nastíněn prvotní impuls pro vznik oboru, spolu s definicí cílů. V druhé kapitole vysvětluji přechod k paradigmatu, v rámci kterého se obor pohybuje dodnes. Také popisuji vznik a fungování perceptronu, tak jak ho nadesignoval Frank Rosenblatt. Tato kapitola také funguje jako popis rozhodovacího procesu, na základě kterého fungují i moderní systémy umělé inteligence. V další kapitole ukazuji, jak za pomoci obohacení perceptronu o další vrstvu neuronů rapidně stoupnou jeho rozhodovací schopnosti. Tento bod je zároveň klíčový pro budoucí vývoj neuronových sítí. Následující kapitola hovoří o konceptu jazykových modelů. Je v ní uvedeno několik přístupů, za pomoci kterých je možné tvořit statistické modely jazyka. Pátá kapitola této sekce obsahuje popis transformerové architektury. Poslední kapitola této sekce, je věnována konceptu před-trénovaných transformerů a konceptu finetuningu.

Sekce B je věnována Wittgensteinovu dílu a především pak jeho pozdní filosofii. V první kapitole této sekce popisuji, jakým způsobem se dělí Wittgensteinovo dílo. Dále také obsahuje popis jeho rané filosofie, přičemž navíc uvádím několik způsobů, jakým jej lze interpretovat. Ke konci kapitoly také popisuji Wittgensteinův radikální odklon od jeho vlastních raných koncepcí. Je tak znázorněn přesun k myšlenkám obsaženým ve Wittgensteinově pozdní filosofii. V následující kapitole popisuji podobu primárního textu Wittgensteinova pozdního období Filosofická zkoumání. Přidržuji se při tom úvodu do studia tohoto díla od Davida Sterna Wittgenstein's Philosophical investigations: an introduction. Je zde prozkoumána jazyková stránka díla spolu s možnostmi, jakými toto dílo lze číst. Tato kapitola také obsahuje krátký popis odlišnosti edice knihy Filosofická zkoumání, se kterou jsem pracoval, od edicí předchozích. Následující kapitola a její podkapitoly jsou věnované několika stěžejním konceptům objevujících se v textu

Filosofická zkoumání. První podkapitola se zabývá kritikou koncepce jazyka sv. Augustina, tak jak ji uvádí Wittgenstein ve svém textu. Zároveň se také jedná o vymezení se Wittgensteina proti svým raným názorům. Druhá podkapitola se zabývá konceptem řečových her. Třetí podkapitola pojednává o konceptu rodinné podobnosti. Čtvrtá podkapitola vykládá Wittgensteinův argument o paradoxu následování pravidla. Pátá a poslední podkapitola uvádí Wittgensteinovy myšlenky ohledně možnosti existence soukromého jazyka a paradoxu, který z něj plyne.

Sekce C je koncipována jako vyvrcholení konceptů, které jsem představil skrze celou práci. Jsou tak přivedené dohromady myšlenky týkající se možnosti modelování přirozeného jazyka a koncepty Wittgensteinovy pozdní filosofie. První kapitola této sekce mluví o významu jazykových modelů. Metody strojového učení jsou zde prozkoumány jako nástroj pro získání nového druhu poznání. Uvádím zde argument o přínosu komputační analýzy, jakožto metody pro studium textu. Je zde uvedený také příklad aplikace, jenž proběhl v minulých letech. Dále je zde zmíněna možnost finetuningu před-trénovaného generativního jazykového modelu na textu lidských autorů. Jsou zde také zmíněny dvě instance, kdy se výzkumníci o tento projekt pokusili. Poslední část této kapitoly je věnována kritické reflexi smysluplnosti dříve zmíněného modelování. Tato kritika plyne s konferenčního příspěvku On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Další kapitola a její podkapitoly slouží jako popis mého pokusu o finetuning před-trénovaného jazykového modelu na primárním textu Wittgensteinovy pozdní filosofie. V první podkapitole uvádím popis mnou zvoleného modelu. Také jsou zde zmíněné mé pohnutky pro volbu tohoto modelu. V další podkapitole popisuji svá trénovací data, za pomoci kterých chci trénovat zmíněný jazykový model. Je zde popsaný proces editace tohoto textu, aby byl vhodný pro strojové zpracování. Jsou zde nastíněny i možné nedostatky tohoto přístupu. V následující podkapitole je popsáno, jak jsem pracoval s před-trénovaným jazykovým modelem. Tato kapitola zmiňuje využití české gridové infrastruktury, MetaCentra. Je zde také uvedeno, jakým způsobem jsem vybíral software pro práci s jazykovým modelem.

1. Popis oboru

Sen o vytvoření inteligentního stroje, který by byl stejně chytrý, nebo i chytřejší než člověk, je starý již několik desítek let. Součástí moderní vědy se stal však až za pomoci digitálních počítačů. Myšlenky, které vedly ke vzniku prvních programovatelných počítačů, vycházejí z pokusů matematiků a logiků o pochopení lidského myšlení a jazyka. Jejich výchozím konceptem byl mechanický proces manipulace se symboly. Digitální počítače jsou ve své podstatě manipulátory symbolů, které jsou schopné provést několik miliard operací za vteřinu (alespoň ty moderní). Pro průkopníky oboru, jako byli Alan Turing a John von Neumann, zde existovaly silné analogie mezi strojem a lidským mozkem. [1]

Počátky oboru AI můžeme vysledovat k workshopu, který probíhal po dobu dvou měsíců v roce 1956 na univerzitě v Darthmouth. Tohoto workshopu se účastnilo deset lidí včetně jeho iniciátora Johna McCarthyho. Cílem workshopu bylo prozkoumání možností strojové inteligence a k ní pojících se témat a technologií jako jsou neuronové sítě, zpracování přirozeného jazyka, formy abstrakce, náhodnost, kreativita a sebezdokonalování. Z nichž všechna tato témata jsou stěžejní součástí oboru dodnes. V návrhu na výzkumný projekt McCarthy se svými kolegy tvrdí následující: „The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer“. Navzdory optimismu čišícímu z celého projektu bylo nejspíš jeho největším přínosem vzájemné seznámení z dnešního pohledu „velikánů“ oboru AI a také vymezení cílů AI jakožto oboru.[2]

2. Sub-symbolické paradigma a Perceptron

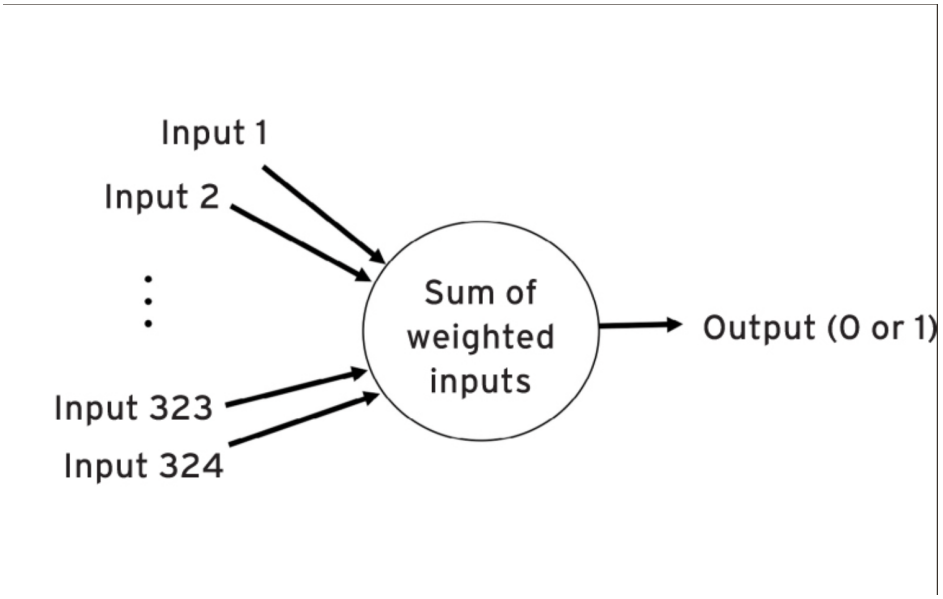
Počátek sub-symbolického paradigmatu byl zapříčiněný především pokrokem v oboru neurovědy. Cílem tohoto přístupu je zachycení nevědomých myšlenkových procesů, které leží pod systémem nazývaným rychlá percepce. Tento systém je mimo jiné odpovědný za rozpoznávání obličejů a identifikování mluvené řeči. Jeho hlavním znakem oproti logicko-symbolickému paradigmatu je absence pevně stanovených překladových pravidel. Namísto nich je stavěno na matematických operacích, které jsou mnohdy hůře

interpretovatelné, než jejich symbolický protiklad. Systémy designované v rámci toho paradigmatu jsou však uzpůsobené k tomu, aby se jim zadaný úkol naučily provádět samy. A to pouze na základě jim nastaveného algoritmu a přeložených vstupních dat. Vzorec jejich úsudku jim tedy není napevno zakódován, ale místo toho emerguje na základě rozpoznaných vztahů. To je motivováno poznatky neurovědců, kteří jsou toho názoru, že právě takto probíhá učení v lidském mozku. Prvotním rozpoznáním vzorce a jeho následným utvrzováním a zpřesňováním při opakovaném výskytu.[1]

Své podněty pro výzkum strojového vnímání uvádí Frank Rosenblatt ve výzkumné zprávě z roku 1957. V té konstatuje, že od nástupu elektronických počítačů, či servo motory poháněných systémů, zde byla fascinace možností sestrojit systém, který by disponoval schopnostmi tradičně připisovanými lidem. Schopnostmi jako percepce, rozpoznávání, formování konceptů a schopností zobecňování na základě zkušenosti. Obzvláštní zájem je pak směřovaný na představu stroje schopného konceptualizace. A to přímo na základě vstupů(světlo, zvuk, teplota, atd.) z fyzického neboli fenomenálního světa. Rosenblatt na základě těchto pohnutek navrhuje systém, jenž by měl být schopný těchto uvedených schopností dosáhnout. A to i bez zásahu jiného agenta (jako například člověka). Toto je hlavním znakem odlišujícím systém od jeho předchůdců. Primárně je od takového systému vyžadováno, aby byl schopný rozpoznat komplexní vzorce informací, které jsou si podobné nebo je spojuje zkušenost. Čili aby byl schopný procesu, který koresponduje s fyziologickými procesy asociace a zobecňování stimulů. Systém by měl tedy být schopný rozpoznat ten stejný objekt nezávisle na orientaci, velikosti, barvě a prostředí, ve kterém se nachází. Tedy aby systém disponoval schopností abstrakce. Zmiňovaný systém operuje na pravděpodobnostním principu místo deterministického. To znamená, že na místo pevně zakódovaných pravidel „pokud, tehdy“ svoji funkci a spolehlivost odvozuje z vlastností statistických operací, provedených na velkém množství elementů. Tento systém, nebo přesněji algoritmus odpovědný za jeho chování, nazývá Rosenblatt **perceptron,** na základě jeho schopnosti percepce. Rosenblatt navrhuje hned několik variant perceptronu dle druhu vstupů, které přijímá. Ve zprávě se však zaměřuje především na model perceptronu, který operuje na základě vizuálních vstupů, tedy fotoperceptronu.[3]

Jak už bylo zmíněno, perceptron svoje fungování staví silně na poznatcích neurověd. Konkrétně na poznatcích o fungování mozku a procesu učení. Neuron je mozková buňka, která přijímá elektrochemické informace (vstupy) z jiných neuronů. Pokud součet všech do neuronu mířících vstupů dosáhne určitého prahu, dojde k jeho „aktivaci“ a neuron pošle

signál dál. Důležité je, že jednotlivé synapse (spojení dvou neuronů) se liší silou propojení. To ve výsledku znamená, že neuron přijímající jejich signály jim přikládá jinou váhu (co se důležitosti týče), než jiným neuronům. Čím větší váha, tím je signálu přidávána větší důležitost. Právě tato charakteristika je dle neurovědců odpovědná za proces učení v lidském mozku. Rosenblatt si však propůjčuje pouze principy těchto procesů a jejich funkci. Jeho snahou není vytvoření kopie mozku. Perceptron je tedy tvořen pouze **umělým neuronem**, ne biologickým, na který jsou připojené vstupy a výstupy. Je možné si tento neuron představit čistě jako abstraktní matematickou funkci odpovědnou za operaci se vstupem. Umělý neuron tedy po vzoru neuronu skutečného, sečte všechny do něj vedoucí signály a pokud výsledek dosahuje předem nastavené prahové hodnoty, pošle odpovídající signál na výstup. Každý vstup je však ještě před sečtením vynásobený přidělenou vahou. Prahová hodnota perceptronu je buďto jednoduše přidělená programátorem, anebo se jí perceptron naučí sám. Výstup může představovat světelná signalizace, nebo jakékoliv jiné zobrazovací zařízení, které nám předá informaci o „výsledku“. Vstup perceptronu se bude lišit dle úlohy, pro jakou se jej rozhodneme nasadit. Pokud po něm budeme vyžadovat zpracování vizuálního úkolu, bude jeho vstupem optický senzor. Umělý neuron můžeme pro naše účely chápat jakožto „black box“, u kterého nic netušíme o jeho vnitřních procesech a jeho vnitřní konstrukci. Důležitá je zde pouze jeho funkce. Ta spočívá v procesu učení-se generovat stejný výstupní signál (nebo tisknout stejné slovo) pro všechny optické podněty, které patří do určité arbitrárně vytvořené třídy. Tedy jinými slovy klasifikování předkládaných vstupů. Dejme tomu, že bychom perceptronu předložili úlohu na rozdělení geometrických útvarů do tříd: trojúhelník, čtverec a kruh. Útvary, které bychom předkládali senzoru, by reprezentovali informační vstup fotoperceptronu. Ten by se mezi nimi učil rozpoznávat jednotlivé odlišující a společné charakteristiky. Následně by pak byl schopný generovat výstup pro každý nově předložený útvar, čímž by nám poskytoval výstupní informaci. Po dokončení trénování by také měl být schopný své nabyté „poznatky“ aplikovat na kterýkoliv jiný dataset, na kterém nebyl trénovaný. Na základě úspěšnosti na nových datasetech se následně určuje „přesnost“ perceptronu.[1, 3, 4]



(schéma perceptronu, obrázek převzatý z[1])

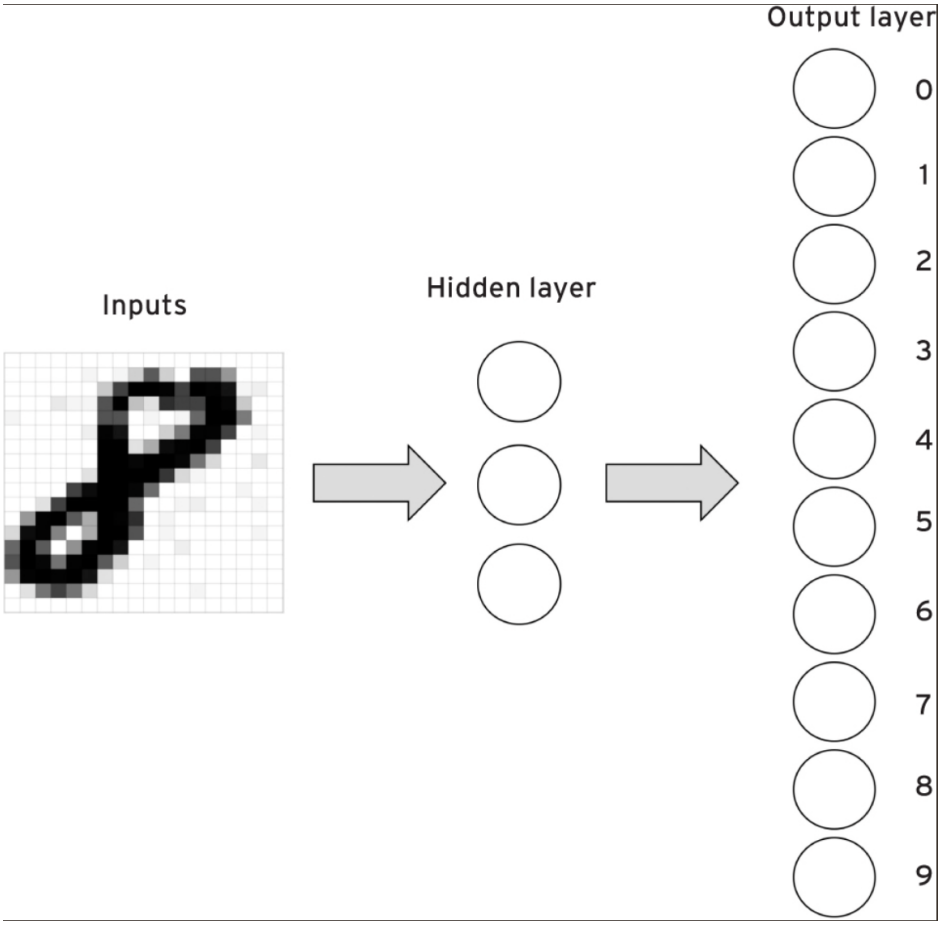
Proces učení probíhá následovně. Dejme tomu, že chceme perceptron naučit rozpoznávat ručně psané číslice od 0 do 9. Číslice budeme předkládat před optický senzor, který je schopný rozpoznat obraz v rozlišení 18x18 pixelů, tedy dohromady 324 pixelů. Toto číslo tak reprezentuje náš potřebný počet vstupů vedoucích do umělého neuronu, tedy 324 vstupů. Pokud bychom operovali s monochromatickým obrázkem, nulová hodnota na vstupu bude znamenat bílou barvu a jedničková hodnota černou barvu. Pokud by nám záleželo i na barvě obrázku, bylo by možné vstupy namapovat na RGB (popřípadě jinou) reprezentaci barev. To by však celý proces notně zkomplikovalo. Pokud bychom chtěli perceptron naučit rozpoznávat ručně psané číslice, znamenalo by to, že potřebujeme 10 výstupních neuronů. Jeden pro každé číslo. Pokud bude jeden z výstupů signalizovat 1, jakožto přítomný signál, znamená to, že perceptron detekoval číslo korespondující s výstupem. Dále zbývá už jen správně rozložení vah jednotlivých vstupů. Rosenblatt na tuto otázku navrhl odpověď inspirovanou behaviorální psychologií. Perceptron by se měl tyto váhy naučit sám. Inspirace pochází z koncepce a pokusů B.F. Skinnera, který trénoval holuby a krysy za použití pozitivní a negativní odměny. Rosenblatt navrhl, že by perceptron měl být trénovaný na příkladech a následně odměněný za úspěchy a potrestaný za chyby. Tato metoda je nyní známá jako **supervised learning**. Ta většinou vyžaduje velké množství trénovacích dat, které obsahují kladné i záporné příklady, které jsou označené. Na začátku učícího algoritmu proběhne náhodné rozdělení vah v rozsahu –1 a 1. Následně je pak perceptronu ukázaný první příklad z trénovacích dat. Perceptron přečte vstupní data a vydá výstup. Pokud je odhad perceptronu shodný s anotací dat, váhy zůstanou stejné. Pokud se výstup liší, perceptron provede malou úpravu vah a

pokračuje k dalšímu příkladu. Perceptron takto projde trénovací data několikrát. Skinner při svých trénovacích experimentech tvrdil, že je lepší provést několik pokusů a postupovat po malých krocích. Může se tak předejít tomu, že by se subjekt naučil nesprávné věci a že by úkol příliš zobecnil. Tento přístup je aplikovatelný i na strojové učení. Může se totiž stát, že algoritmus dosáhne až moc obecného natrénování a nebude pak fungovat na konkrétních příkladech. V momentě, kdy je algoritmus úspěšný na všech trénovacích datech, je učení zastaveno a je možné se přesunout k testovacím datům. Na těch se ukáže, jak si perceptron vede na datech na kterých nebyl trénován a jak moc je použitelný ve „skutečné" aplikaci.[1, 4]

Marvin Minsky a jeho kolega z Massachusetts Institute of Technology, Seymour Papert v roce 1969 publikovali knihu nazvanou *Perceptrons, ve které* ukazují, že schopnosti perceptronu jsou značně omezené a že se jeho využití hodí pouze pro určité úkoly. Další překážkou pro perceptron by byly úkoly s požadavkem na větší množství vah. To kvůli způsobu jakým funguje učící algoritmus Perceptronu. Minsky s Papertem však dodávají, že přidáním už jedné skryté vrstvy neuronů, by se schopnosti perceptronu značně rozšířili a mohl by tak být použitý při řešení rozličnějších úloh. Perceptron s takovouto skrytou vrstvou se pak nazývá **vícevrstvá neuronová síť**. Tyto sítě pak tvoří základ pro moderní AI.[5]

3. Vícevrstvé neuronové sítě

Neuronová síť můžeme označit jakožto vícevrstvou, pokud je sestavená z dvou a více vrstev neuronů. Tyto další vrstvy se tradičně nazývají skryté, což jednoduše znamená „ne-výstupní“. Vícevrstvá neuronová síť tedy může mít libovolný počet skrytých vrstev a jednu výstupní vrstvu. Většinou je však těžké odhadnout kolik skrytých vrstev by mělo být nasazeno a kolik neuronových jednotek by měla každá skrytá vrstva obsahovat. Výzkumníci zabývající se studiem neuronových sítí většinou volí metodu pokus–omyl, aby dosáhli nejlepší konfigurace topologie. V jiných instancích většinou výzkumníci adoptují strategii jiného týmu, který se podílel na podobném projektu. V případě vývoje transformeru GPT-NeoX se jeho tvůrci rozhodli převzít většinu designu z GPT-3. [1, 6]



(schéma vícevrstvé neuronové sítě, obrázek převzatý z[1])

Podobně jako u perceptronu má každá neuronová jednotka za úkol vynásobit své vstupy jejich příslušnou vahou a následně sečíst jejich výsledky. Nicméně výstup neuronové jednotky v tomto případě už není tak jednoduchý jako u perceptronu. Nejedná se totiž o prosté stavy na základě prahové hodnoty, kdy je výstup aktivní, nebo ne (1 a 0). Místo toho využívá každá jednotka součet vstupů k výpočtu čísla na škále mezi 0 a 1, které se nazývá jeho **aktivace**. Pokud je vypočtená hodnota nízká, bude se aktivace jednotky blížit k 0. Pokud bude vypočtená hodnota vysoká, bude se aktivace blížit k 1. Za účelem rozpoznání ručně psaného čísla (jak tomu bylo v kapitole pojednávající o perceptronu) provádí síť výpočty po jednotlivých vrstvách z levé strany sítě do pravé. Každá skrytá vrstva vždy předává svou vypočítanou aktivační hodnotu. Z těch se následně stávají vstupy pro další vrstvu. Tento proces pokračuje dokud se výpočet aktivace nedostane až k výstupní vrstvě, kde je předaný na výstup (zobrazovací zařízení). Pokud bychom chtěli vícevrstvou síť použít pro rozpoznávání čísel jakožto v předchozím příkladu, měli bychom stále deset výstupů. Na nich bychom však už nedostávali signál detekováno / nedetekováno (1 a 0), ale místo toho samotné aktivační hodnoty, které mají , jak už bylo řečeno, podobu desetinného čísla mezi 0 a 1. Je tedy možné je chápat jakožto míru

(procento) toho, jak moc si je síť „jistá“ výskytem dané číslice. Výstup s nejvyšší aktivační funkcí můžeme pak chápat jakožto výslednou klasifikaci vstupu, neboli „odpověď“ sítě.[1]

V principu se dá vícevrstvá neuronová síť natrénovat k rozpoznávání mnohem abstraktnějších **znaků** (features), než jsou jednoduché pixely. Například vizuální tvary, jako jsou obloučky v ručně psané číslici 8. Je však dobré uvést, že oproti dříve zmiňovanému perceptronu, tento druh neuronových sítí vyžaduje jiný algoritmus nazývaný **back-propagation** (zpětné šíření). Jak už název napovídá, jedná se o algoritmus, který detekuje chybný výstup, dejme tomu špatně rozpoznané číslo, a zpětně vysleduje, jaké propojení sítě může za danou chybu. To umožňuje identifikovat špatně rozdělené váhy sítě a jejich následnou korekci vůči očekávanému výstupu. Rozporu mezi reálným a očekávaným výstupem se říká **ztrátová funkce** (loss function). K jejímu výpočtu můžeme použít několik dostupných metod. Volba záleží především na typu problému k jehož řešení plánujeme neuronovou síť použít. Při zpětném průchodu sítí vypočítá algoritmus gradient ztrátové funkce s ohledem na každý parametr sítě. Tyto gradienty poskytují informaci, jakou mírou ovlivní změna určitého parametru ztrátovou funkci. Tím je možné určit, jaké parametry je třeba změnit za účelem zmírnění ztrátové funkce. Po výpočtu ztrátové funkce a gradientů přijde na řadu algoritmus **gradient descent** (postupného sestupu). Jeho cílem je optimalizace parametrů sítě. Tento algoritmus iterativně (každá iterace se nazývá epochou) upravuje parametry sítě proti směru gradientů tak, aby se docílilo minima ztrátové funkce. Parametry jsou upravovány odečtením zlomku gradientu od momentální hodnoty parametru. Tento zlomek je možné zvolit manuálně a reprezentuje takzvanou **rychlost učení** (learning rate). Ta reprezentuje, jak velké kroky budou v rámci iterací probíhat. Čím větší rychlost učení, tím bude algoritmus podnikat větší kroky ve směru nejvyššího poklesu ztrátové funkce. Menší rychlost učení má za následek pomalejší dosažení minimální ztrátové funkce, ale na druhou stranu poskytuje větší stabilitu sítě. Hrozí tak menší riziko, že se síť naučí rozpoznávat špatné symboly. Větší rychlost učení má za následek pravý opak, rychlé dosažení minima a menší stabilitu. Nalezení správné rychlosti učení je tak jedním z důležitých aspektů při procesu trénování.[1, 7]

Jak už bylo řečeno, vícevrstvé neuronové sítě mohou rozpoznávat mnohem abstraktnější znaky než perceptron. S tím je úzce spojený koncept designování skrytých vrstev. Zůstaňme u příkladu s vícevrstvou sítí, která má za úkol rozpoznávat číslice od 0 do 9. Designér této sítě by mohl předpokládat následující. Tyto číslice jsou sice na první pohled odlišné, každá z nich je ale složená z podobných znaků. Pokud například porovnáme

číslice „8“ a „9“, můžeme mezi nimi vidět podobnosti. Obě číslice jsou zcela tvořené obloučky a některé jsou použité v případě obou číslic. Bylo by dokonce možné definovat číslici „8“ jakožto „soubor znaků, tvořící číslici 9, který navíc obsahuje levý dolní oblouček“. Můžeme se tak pokusit kolem těchto znaků vytvořit design sítě. Dejme tomu, že identifikujeme sadu šesti unikátních znaků, ze kterých jsme schopni vytvořit všechny číslice. Znamenalo by to tak, že naše skrytá vrstva by se skládala ze šesti neuronových jednotek. Na základě rozpoznání znaků v této vrstvě by pak výstupní vrstva obdržela informaci, jaké znaky byly předchozí vrstvou rozpoznány, a provedla vyhodnocení, o jakou číslici se jedná. Je pak možné si představit, že bychom podobným způsobem přidávali další vrstvy za účelem zjemnění procesu rozhodování. Takto by bylo možné rozložit jemné znaky, které rozpoznáváme v číslicích na ještě menší části. V teorii by se tak dalo dosáhnout větší rozpoznávací schopnosti sítě a nuance. Je však nutné uvést, že není zaručené, že se síť během trénování naučí rozhodovat právě na základě těchto námi zamýšlených znaků. Během trénování můžeme síti signalizovat pouze, jaký výstup očekáváme a zda klasifikace proběhla úspěšně. Jaké znaky uvnitř sítě budou rozpoznány a jak budou ovlivňovat následující vrstvy, můžeme označit téměř za náhodné. Nemáme také možnost určit, jaký neuron je odpovědný za rozpoznání jakého znaku. Toto podobenství je pouze interpretací rozhodovacího procesu neuronové sítě. Nepochybné je však to, že přidáváním vrstev se zvyšuje rozlišovací jemnost sítě.[7]

Tato myšlenka přidávání vrstev vedla až k vytvoření **hlubokých neuronových sítí** (deep neural networks). Jedná se o podkategorii vícevrstvých neuronových sítí. Není daná rigidní hranice, kdy se síť může začít nazývat hlubokou. Co tento název však označuje, je stav při kterém neuronová síť naroste do takové velikosti, kdy už nemůžeme s jistotou říci, co se děje ve skrytých vrstvách sítě. Jinými slovy, rozpoznávané znaky se stanou tak jemnými, že je nedokážeme identifikovat. To však pro tyto sítě nebylo překážkou, aby nabyly na popularitě. Naskýtá se zde však prostor pro debatu ohledně etických implikací využití těchto sítí. Můžeme si klást otázku, zda je vhodné nasazovat tyto sítě a brát v potaz jejich výstupy, pokud nevíme, čím bylo ono rozhodnutí ovlivněno. Celý systém se pro nás najednou stává „black boxem“, ve kterém se otevírá prostor pro všelijaké biasy. Právě z tohoto důvodu je v posledních letech kladený důraz na takzvanou Explainable artificial intelligence (vysvětlitelnou umělou inteligenci), která si bere za úkol tyto nedostatky napravit.[1, 7]

4. Jazykové modely

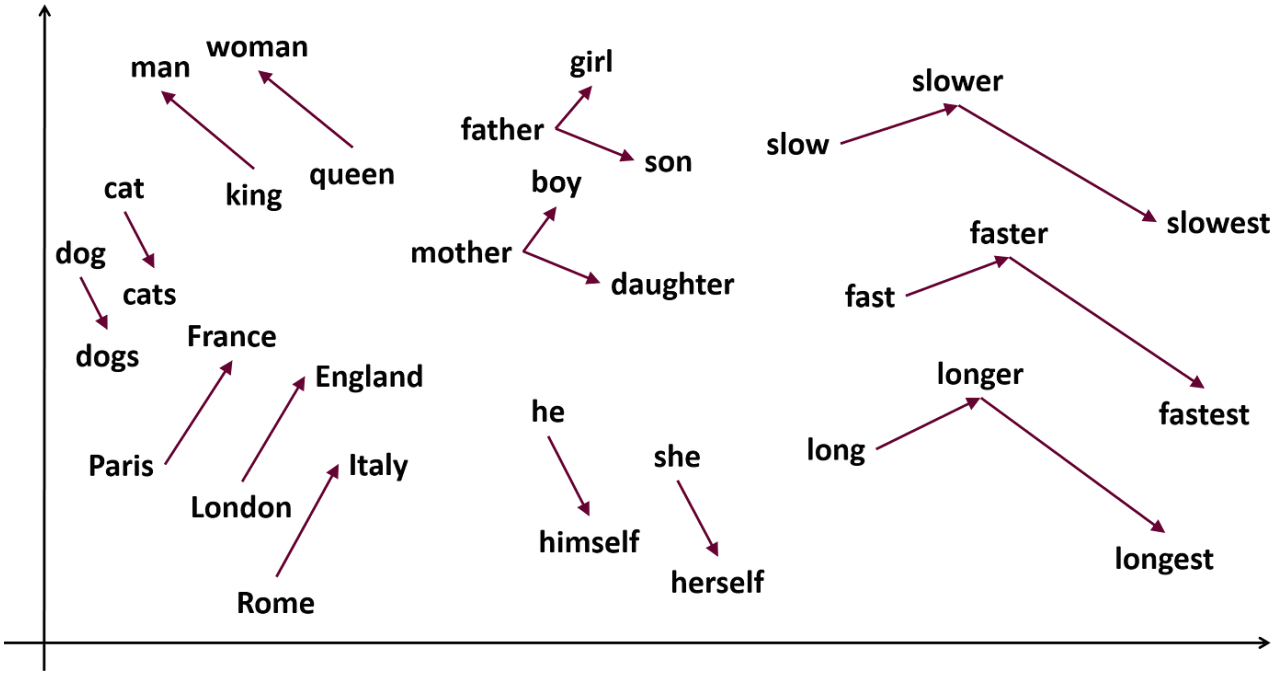
Jazykový model je termín používaný v oboru NLP, který označuje pravděpodobnostní rozdělení pro danou sekvenci slov. Jinak řečeno, pravděpodobnost, že se daná sekvence slov vyskytne v daném kontextu. Slovem model se zde tedy míní jakýsi obraz, nákres, či nápodoba toho, jak s jazykem zacházíme v našem běžném životě. Učením se na velkém množství textu, jsou tyto modely schopné zachytit spletitosti gramatiky, sémantiky a syntaxe. Jakožto hlavní přístupy k modelování jazyka můžeme označit **pravděpodobnostní jazykové modely** a **modely založené na vektorové sémantice**.[1, 7]

Při pravděpodobnostním přístupu je modelovaná sekvence slov často označovaná jako **n-gram**, tedy sekvence *n* slov. Sekvenci dvou slov bychom nazývali *bigram*, sekvenci tří slov *trigram*. Příkladem bigramu mohou být sekvence „dobrý den“, „bílá stěna“, „známá tvář“. Formálně zapsaný n-gram pak vypadá takto: $P(w|h)$. Tento zápis nám říká, jaká je pravděpodobnost *P*, že se slovo *w* vyskytne v závislosti na historii *h*. Tento zápis tedy můžeme naplnit větou „Dnes je venku slunečno“: $P(slunečno|dnesjevenku)$. *P* nám v tomto případě označuje, jak pravděpodobné je, že sekvence *„dnes je venku“* bude pokračovat slovem „*slunečno*“. Teoreticky, tato pravděpodobnost následujícího slova se dá vypočítat jednoduše. Pokud bychom měli dostatečně velký korpus, stačilo by spočítat kolikrát je daná sekvence následovaná konkrétním slovem. V praxi však nastává ten problém, že ve spoustě případů nám ani *world wide web* neposkytne dostatečně velký korpus, abychom dospěli k dobrému odhadu. To z toho důvodu, že jazyk je fluidní věc, která se neustále vyvíjí. Vznikají nové způsoby použití slov a místy dokonce i nová slova. K tomu je však možné tato slova poskládat v nekonečném množství kombinací. Je tedy možné, že se i poměrně jednoduché (a běžné) sekvence slov v korpusu nemusí vyskytovat ani jednou. Tento problém lze ale vyřešit za pomoci aproximace. Namísto toho, aby se pravděpodobnost počítala na základě celé historie slov, je jí možné aproximovat za pomoci několika posledních slov. Tento přístup je založený na řetězovém pravidlu pravděpodobnosti, dle kterého je možné chápat pravděpodobnost jedné sekvence slov, jakožto součin všech jejích částí. Je tak možné namísto pravděpodobnosti n-gramu: $P(postel|nadruhéstraněpokojestálapřekvapivěveliká)$, spočítat pouze pravděpodobnost pro bigram: $P(postel|veliká)$. Tedy za předpokladu že tuto pravděpodobnost následně vynásobíme předcházejícími částmi sekvence. Výsledná pravděpodobnost celého n-gramu je tedy rovná součinu jednotlivých po sobě jdoucích bigramů, ze kterých je

celkový n-gram složený. Tomuto přístupu se také říká Markovův řetězec, což je stochastický model popisující pravděpodobnost přechodu z jednoho stavu do druhého, pouze na základě předchozího stavu. Takto můžeme bigram zobecnit na trigram, a ten zobecnit na libovolný n-gram. [7, 8]

Dalším přístupem k modelování jazyka, je **vektorová sémantika**, která bývá někdy také označována jako distribuční sémantika. Hlavní myšlenkou tohoto přístupu, je reprezentace slov za využití vektorů v mnohodimenzionálním prostoru. Slova jsou reprezentována hustými číselnými vektory, kdy každá dimenze vektoru koresponduje s určitým znakem (feature) či atributem. Ty souvisejí se způsobem použití slova nebo s jeho kontextem. Vektory jsou obvykle mnohodimenzionální a spojité, což umožňuje vyjádřit jemnější nuance použití (významu) slova. Klíčová je zde distribuční hypotéza, podle které je možné zachytit význam slova analýzou jeho distribučních vlastností uvnitř korpusu. Ta předpokládá, že slova stejného nebo podobného významu mají tendenci se vyskytovat ve stejných kontextech. Analyzováním vzorců společného výskytu slov ve velkém korpusu se vektorová sémantika snaží zachytit významové vztahy a podobnosti mezi slovy. Vektorová reprezentace slova je označována jakožto **word embedding**[1]. Existuje několik metod, za pomoci kterých je možné dosáhnout vektorové reprezentace slov (např: tf-idf, word2vec). Důležité je však zmínit, že každá z nich využívá jiných metod, jak pojmout slova vyskytující se v korpusu. To má za následek, že informace získaná analýzou textu se liší na základě nasazené metody. Jako vždy má každá metoda své biasy a z nich plynoucí klady a zápory. Je tak nutné, jako vždy, zvolit správné nástroje pro úkol, se kterým se chceme potýkat. [7]

[1] Raději se přikláním k anglickému originálu, než k české obdobě termínu „vnořené slovo“



(Word2vec reprezentace slov, obrázek převzatý z[9])

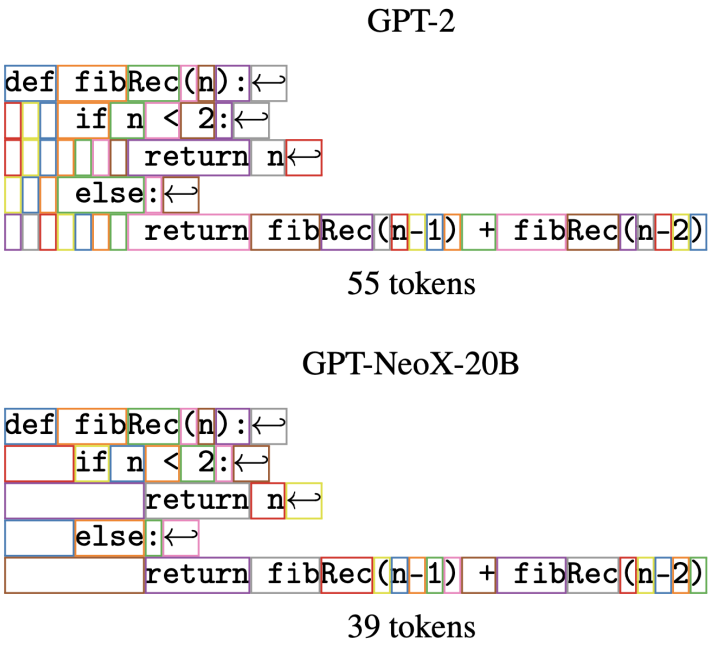
Důležitý konceptem v posledních letech jsou takzvané **velké jazykové modely** (large language models). Jejich hlavní charakteristikou je návodně velikost jejich parametrů. Parametry zde znamenají počet nastavitelných hodnot, který je model schopný se naučit během procesu trénování (word embeddingy, váhy neuronové sítě, váhy attention mechanismu). Ty v případě velkých jazykových modelů dosahují běžně hodnot v řádu milionů a v období posledních let často i miliard (za příklad můžou sloužit modely GPT-4, BERT a Llama). Tyto modely vznikají za využití obrovského množství energie a výpočetní síly. Bývají trénované na rozsáhlých textových datasetech získaných scrapováním (stahováním obsahu z webových stránek, většinou ve formě přirozeného jazyka) internetu. Nutné je však dodat, že ne vždy větší množství parametrů znamená větší sofistikovanost modelu.[6, 10–12]

5. Transformery

Transformery byly poprvé představeny v roce 2017 ve článku *Attention Is All You Need*. Od té doby se díky nim spustila nová vlna zájmu o NLP a obor umělé inteligence. Tato architektura zatím posloužila jakožto základ hned několika „*state of the art*“ jazykových modelů. Jejich nasazení vedlo k velikým pokrokům v oborech strojového překladu, sentiment analysis (analýzy sentimentu), systémech pro zodpovídání otázek a generování textu. Oproti starším modelům jako jsou rekurentní neuronové sítě (RNN), které zpracovávají vstup sekvenčně (popořadě), jsou transformery výkonnější a efektivnější, co

se úloh zpracování přirozeného jazyka týče. Toho jsou schopné za využití attention mechanismu, který jim umožňuje zpracovávat vstupy (slova, tokeny) paralelně. Díky tomu jsou transformery schopné rozpoznávat i vzdálené[2] závislosti, což je činí schopné zachytit komplexní lingvistické vzorce a kontexty.[7, 13]

Transformerovou architekturu si můžeme představit na základě následujícího příkladu. Jedná se pouze o ilustrační popis na základe článku *Attention Is All You Need* (V příloze na konci práce je poskytnutý diagram transformeru, který popisuji v následující části). V jádru transformerových modelů stojí dva fundamentální komponenty, enkodér (levá strana diagramu) a dekodér (pravá strana diagramu). Enkodér přijímá sekvenci vstupních tokenů, což jsou základní jednotky vstupního textu, které algoritmus zpracovává. Token může mít formu jednoho znaku, podslova, nebo slova.



(ukázka tokenizace, obrázek převzatý ze[6])

Token vzniká za použití procesu tokenizace, při kterém je text rozložený na tyto menší jednotky. Velikost tokenu, pak podléhá algoritmu zvolenému pro jeho získání. Přesný popis fungování algoritmů tokenizace a k němu pojících se vlastností je nad rámec této práce. Důležité je však zmínit, že se pohybujeme v odvětví strojového učení, které často své statistické operace neprovádí na prostých slovech, nýbrž na podjednotkách, které slova tvoří. Tato vstupní sekvence tokenů je následně vložena do enkodéru, který se skládá z několika vrstev. Každá vrstva obsahuje self-attention mechanismus a feed-forward neural

[2] Vzdáleností je myšlena délka textu mezi dvěma slovy.

network (dopřednou neuronovou síť). Vstupní sekvence je nejprve převedena do hustých vektorových reprezentací[3] či embeddingů v kontinuálním prostoru. Embeddingy jsou obvykle vytvářené náhodně a v průběhu procesu učení upravované tak, aby smysluplně reprezentovaly tokeny. Self-attention mechanismus následně umožňuje transformeru zhodnotit důležitost tokenu v rámci celé vstupní sekvence. V něm je každý embedding převeden na tři vektory Query (dotaz), Key (klíč) a Value (hodnotu). Ty jsou poté použité k výpočtu attention skóre. To je vypočítáno pro každý token vůči všem jednotlivým tokenům ve vstupní sekvenci. Na základě tohoto skóre systém určuje, jak moc jsou jednotlivé tokeny závislé na ostatních tokenech. Jinak řečeno, jakým tokenům by měl systém „věnovat pozornost“. Tokeny s větší attention vahou mají větší vliv na finální reprezentaci každého tokenu. Transformer však tyto attention váhy vypočítá hned několikrát. Každá self-attention část má totiž několik „hlavic“ (heads), které nezávisle na sobě vytvoří attention skóre. Tyto self-attention části se proto nazývají „multi-headed attention“. Výsledné attention skóre se následně vypočítá na základě vah z všech hlavic. Po jeho získání je výstup poslán skrze **feed-forward neural network**. Tento druh neuronové sítě je specifický tím, že je vícevrstvý a informace v něm proudí jen jedním směrem (tradičně kresleno zleva doprava). Konceptuálně se tak neliší od příkladu vícevrstvé neuronové sítě, který byl uvedený dříve. Feed-forward sít následně na získané reprezentace aplikuje nelineární transformaci. Vstupní sekvence takto projde několika vrstvami encodéru, přičemž je její reprezentace takto aktualizována v každé vrstvě. Po průchodu každou vrstvou jsou aplikována „residual connections“ (zbytková spojení) a následně normalizace vektorů. Residual connections umožňují transformeru předávat původní reprezentace tokenů přímo do další vrstvy, čímž je zmírňován problém mizejícího gradientu (problém vznikající při učení za využití back-propagation algoritmu). Výstup je také obohacený o poziční kódování, které indikuje pořadí slov. Enkodér je tak ve výsledku odpovědný za zpracování kontextu ve vstupní sekvenci.[7, 13]

Dekodér je sestavený ze stejného počtu vrstev jako enkodér. Jejich složení se však nepatrně liší. Na začátku je totiž každá vrstva dekodéru doplněna o masked multi-head (maskovaný vícehlavý) mechanismus. Poté už design dekodéru kopíruje enkodér. Jakožto vstup dekodér přijímá buďto svůj vlastní výstup z minulé iterace, k čemuž dochází během procesu trénování, nebo dříve vygenerovaný token v průběhu inference („dotazování“ již natrénovaného modelu). Podobně jako u enkodéru se vstupní sekvence dekodéru nejprve

[3] Hustá reprezentace v oboru NLP znamená, že většina prvků uvnitř vektoru je nenulová.

transformuje na hustou vektorovou reprezentaci. Ta slouží dekodéru jakožto počáteční reprezentace vstupní sekvence. Dekodér na začátku taktéž využívá self-attention mechanismus (opět se stejným počtem hlavic jako enkodér), ale s tím rozdílem, že je doplněný o takzvanou „masku“. Ta zapříčiňuje, že self-attention mechanismus věnuje pozornost pouze předcházejícím (dříve generovaným/přečteným) tokenům. Maska totiž zakrývá všechny budoucí tokeny. Dále dekodér využívá cross-attention, za pomoci které přihlíží k výstupu enkodéru během zpracování každého tokenu. To dekodéru umožňuje se zaměřit na relevantní části vstupní sekvence a je tak zajištěno, že výstup dostává kontextuální informaci ze vstupu. Výstup dekodéru je však pouze vektor složený z čísel s pohyblivou řádovou čárkou. Je pak prací dvou posledních vrstev transformeru (Linear a Softmax vrstvy), aby je přeměnily ve slova. Zde v podstatě probíhá výběr nejpravděpodobnějšího tokenu ze všech dostupných možností. [7, 13]

Transformerové modely našly mnoho uplatnění. Jednou z prvních oblastí jejich aplikace byl strojový překlad. Transformer byl natrénovaný na datasetu, který obsahoval text a jeho překlad. Na základě těchto trénovacích dat byl transformer natrénovaný na „které tokeny v jakých kontextech korespondují s jakými tokeny v jejich kontextech“. Jinak řečeno, „jaká věta francouzského jazyka koresponduje s jakou větou jazyka anglického“. Na podobném principu funguje generování „nového“ textu ve stejném jazyce. Jakožto vstup je transformeru předložený řetězec tokenů (před tokenizací věta v přirozeném jazyce) na základě kterého se vygeneruje nejpravděpodobnější pokračování dané sekvence.[7, 14]

Ke trénování transformeru jakožto jazykového modelu se dá použít několika metod, přičemž jednou z nich je **self-supervised learning** (učení se pod vlastním dohledem). Jedná se o metodu, za použití které se systém, v našem případě transformer, naučí rozpoznávat vlastnosti (features) vstupních dat, bez toho aniž by byly předem anotované člověkem. Systém je naopak schopný dohlížet na své trénování sám. Stojí tak na pomezí mezi supervised (učení s učitelem) a unsupervised (učení bez učitele) learning přístupy. V našem případě totiž transformeru postačí původní korpus textu, na kterém je trénovaný. Přirozenou sekvenci slov, vyskytující se v korpusu, můžeme totiž chápat jakožto anotaci dat, jak by měl přirozený jazyk vypadat. Je tomu tak, protože se snažíme zachytit vlastnosti textu, na kterém je model trénovaný. Jednoduše tak trénujeme model k tomu, aby minimalizoval chybu v předpovídání následujícího slova ve trénovací sekvenci. Úprava vah pak probíhá za využití backpropagation algoritmu, jak byl popsán v předchozí kapitole. Trénování může být ukončeno v moment, kdy uznáme že model se naučil vše, co

jsme od něj vyžadovali. K tomu nám mohou dopomoci jednotlivé metody určené k vyhodnocení natrénovaného modelu (Perplexity, BLEU, NER, atd.).[7, 15]

Jakmile je model hotový s trénováním, je možné ho použít ke generování nového textu. Při generování model buďto nejprve náhodně vybere slovo vhodné jako začátek sekvence, nebo ho poskytne uživatel takzvaným promptem (text který se bere jakožto start řetězce generování). Následně pak model vybírá slova na základě dříve vygenerovaných slov sekvence. Takto postupuje, dokud není dosaženo předem specifikované délky sekvence nebo není vygenerovaný koncový symbol sekvence. Takový způsob generování se nazývá **autoregresivní**, což znamená, že model předpovídá hodnotu v čase *t* na základě lineární funkce předchozích hodnot v čase *t* – 1 atd. Toto označení se používá i u jazykových modelů, a to i přestože nefungují lineárně. Použití jazykových modelů ke generování textu je jednou z oblastí (spolu s generováním obrazu a počítačového kódu), kde byl dopad neuronových sítí na oblast NLP největší.[7, 15]

6. Před-trénované transformery a finetuning

Jedná se v podstatě o předem natrénované modely, které disponují před-porozuměním úkolu, ke kterému jsou určené. Za příklad nám mohou sloužit jedny z nejznámějších modelů tohoto typu: BERT a GPT. Tyto modely je možné „doladit“ (finetune) na specifické úkoly. Teorie za tímto přístupem je taková, že pokud model bude napřed natrénovaný na obecných datech, dostane se mu obecnému porozumění úkolu, na který bychom ho chtěli nasadit. Poté už jenom stačí tento model doladit na specifických datech, která nás zajímají. Toto dolazení probíhá stejným způsobem, jako původní trénování modelu. Na základě vloženého textu na dolazení a self-supervision metody (nebo jiné dostupné) si model upravuje své váhy. Model si tímto způsobem zachová své obecné porozumění, ale obdrží „biasy“ z dat, na kterých byl dolazený. Tato metoda může být výhodná, pokud jsou naše data velmi malá na to, aby byl model trénovaný čistě na nich. Dalším bodem zájmu může být možnost sdílení před-trénovaného modelu (nezávisle na velikosti). Velké jazykové modely jsou trénované na velikém množství dat a z toho důvodu dosahuje jejich trénovací proces veliké energetické spotřeby. V případě trénování velkého jazykového modelu Llama od společnosti Meta se odhaduje energetická spotřeba na 2,638 MWh. K té je navíc zapotřebí přístup k dostatečně výkonnému hardwaru, který by byl schopný trénovací proces provést v rozumném časovém horizontu. Tvorba takto velkých modelů je tedy už z principu umožněna pouze velkým organizacím. Pokud je tedy model

zpřístupněný a dosahuje dostatečné obecnosti, je možné ho za pomoci procesu dolazení přizpůsobit pro své účely. To celé za zlomek trénovací náročnosti. Tento koncept sdílení již natrénovaných modelů má také pozitivní dopad na ekologickou stránku trénovacího procesu, jelikož hlavní část (pokud se rozhodneme vynechat proces dolazení, tak také jediná) trénování proběhne pouze jednou. Tato koncepce sdílení je aplikovatelná nezávisle na velikosti. Vždy je vhodné myslet na energetický (a v tomto smyslu i ekologický) dopad našeho výzkumu. Pokud tedy tyto myšlenky přivedeme zpět do oblasti modelování přirozeného jazyka, jedná se o přesně o výsledek který potřebujeme. Transformery, které byly před-trénované na velkých datových korpusech, jsou totiž schopné postihnout „vtělené znalosti“ (embodied knowledge) syntaxe a sémantiky přirozeného jazyka. Tímto způsobem je tak v teorii možné vytvořit jazykový model, který bude schopný porozumění obecnému přirozenému jazyku ale zároveň i porozumění specifickému textu. K dolazení transformeru na specifickou úlohu nám navíc stačí jen malý objem dat. To vše má za následek zvýšený výkon modelu, na specifických úlohách.[10–12, 16–18]

Sekce B - Wittgenstein a jeho pozdní filosofie

Tato sekce práce slouží především jakožto úvod do Wittgensteinovi pozdní filosofie. Mým cílem není provedení radikální nové interpretace, či obhájení Wittgensteinových tezí. Jedná se pouze o představení několika hlavních konceptů vyskytujících se v první části díla *Filosofická zkoumání*. Je samozřejmě možné v díle zpozorovat více konceptů, které se dají označit za důležité či inspirativní. Takový rozbor by byl však nad rámec této práce. Koncepty představuji především z důvodu obeznámení čtenáře, který by případně neměl jejich znalost. Tato část tedy primárně slouží jakožto srovnání pro výstupy generované mnou trénovaným jazykovým modelem.

V této sekci se také vyskytuje úvod k Wittgensteinovu filosofii jako celku. Jedná se spíše o zasazení spisu *Filosofická zkoumání* do kontextu v rámci Wittgensteinova díla. *Zkoumání* jsou zde pak uvedena do kontrastu oproti jeho raným názorům na koncepci a fungování jazyka.

Většina interpretace Wittgensteinova textu vyskytující se v této sekci práce je převzatá od Davida Sterna a jeho textu *Wittgenstein's Philosophical investigations: an introduction*. Tento text jsem se rozhodl použít na doporučení vedoucího této diplomové práce. Stern se v tomto textu zabývá především strukturou díla a představení konceptů jež se v něm vyskytují. Zvláštní pozornost věnuje také zdůraznění jazykové stránky textu a tomu jak jsou v díle vystavěné struktury argumentů. Přičemž obojí hraje důležitou roli ve správné interpretaci Wittgensteinových myšlenek.

V této teoretické sekci popisující Wittgensteinovo dílo, pracuji výhradně s anglickou literaturou a anglickými překlady Wittgensteinových textů. To z toho důvodu, že tato práce je zaměřená na natrénování jazykového modelu na anglickém překladu Wittgensteinova textu. Přišlo mi tak logické, abych i já ve svém úvodu do Wittgensteinovi filosofie pracoval s tímto jazykem. Z tohoto důvodu jsou veškeré přímé citace v následující sekci v anglickém jazyce.

1. Od Traktátu ke Zkoumání

Tradičně je Wittgensteinovo dílo rozdělované na dvě stěžejní období, rané a pozdní. Ačkoliv tato koncepce dělení může být, a často také je, podrobená kritice, plánuji se jí přidržet při psaní této práce. Jedna z častých kritik tohoto dělení je, že myšlenky, které nacházíme v pozdním období, jsou pozorovatelné již v jeho takzvaném středním období.

Centrální myšlenky první části *Filosofických zkoumání* se u Wittgensteina vyskytovaly již v době, kdy vznikal text dnes známý jako *Brown Book*. Můžeme být dále vznesena námitka, že samotná koncepce rigidního třídění lidského myšlení na jasně dělitelné body nedává smysl. Bylo by snad pro Wittgensteina vůbec možné dospět ke svým pozdějším myšlenkám, nebýt toho, jakým způsobem myslel dříve? Z tohoto hlediska by dávalo větší smysl pojímat Wittgensteinovo dílo, nebo dokonce i dílo jakéhokoliv autora, jakožto jeden fluidní celek. Určité myšlenky, témata, podněty a jejich evoluce v čase. Nicméně je také možné chápat toto dělení, jakožto orientované na dva Wittgensteinovo stěžejní texty *Traktát logicko-filosofický* a *Filosofická zkoumání*. V tomto smyslu pak můžeme chápat tato dvě pomyslná období jakožto body, ve kterých konvergují dva radikálně odlišné pohledy na jazyk, myšlení a jejich vztah ke světu, koncentrované do dvou textů. *Traktát* jakožto suma jeho raného období, *Zkoumání* jako suma období pozdního. Toto dělení je také možné ospravedlnit na základě autorství textu. Přestože se nepochybuje, že autorem textů připisovaných Wittgensteinovi je skutečně on, je zde prostor pro kritiku editorské práce. To z toho důvodu, že za kompilaci textu byl Wittgenstein odpovědný pouze v případě *Traktátu* a první části *Zkoumání*. Další texty, jako je například: *Blue and Brown Book*, *On Certainty* a *Big Typescript* byly ve skutečnosti kompilovány správci Wittgensteinovy pozůstalosti. A přestože, jak už bylo řečeno, zde není pochyb, že autorem jednotlivých myšlenek je Wittgenstein, je možné namítat, že výsledná kompilace nemusí odpovídat autorovu záměru. Své zaměření na Wittgensteinovu pozdní filosofii tedy míním ve smyslu zaměření se na text *Filosofická* zkoumání, konkrétně na jeho první část.

Wittgensteinovo rané období se vyznačuje především idealismem či přesněji, logickým atomismem. Za pomoci moderní logiky v něm propojuje mysl, jazyk a žitý svět a vypovídá tak o vztazích, které mezi těmito sférami existují. Interpretací tohoto někdy až enigmatického textu existuje nespočet, což můžeme ilustrovat na už nechvalně známé sedmé větě. „Whereof one cannot speak, thereof one must be silent“. Někteří autoři (Vídeňský kroužek) jí považovali za odmítnutí metafyziky a tím pádem i jako „náboj“ do jejich programu proti metafyzice. Dle nich je tato věta pejorativní a říká nám, že o metafyzice se nedá logicky hovořit a tudíž bychom s ní neměli ztrácet čas a raději o ní pomlčet. Toto čtení také utvrzoval důraz na rozlišení smysluplného a nesmysluplného, který je v *Traktátu* přítomný. Jiní autoři v díle spatřují pravý opak, který pouze ukazuje schopnosti a limity našeho jazyka. Sedmá věta pak podle nich konstatuje „zde jsem vybudoval vše, co pojme logika/náš jazyk, to ostatní však musíme zažít/ukázat“. V tomto

pojetí by tak sedmá věta nebyla pejorativní. Jednoduše by znamenala, že některé věci jsou nad síly logiky a našeho jazyka a je nutné s nimi přijít do přímého kontaktu. Je dokonce možné v *Traktátu* identifikovat celý mystický aspekt. Sedmé větě pak můžeme připisovat význam, že některé věci vyžadují určité zasvěcení či zkušenost (Jak je uvedeno v předmluvě: „This book will perhaps only be understood by those who have themselves already thought the thoughts which are expressed in it“) a teprve až skrze ně, je pak možné určité věci pochopit. Jak říká věta 5.6: „The limits of my language mean the limits of my world“. V tomto smyslu můžeme o *Traktátu* uvažovat jako o něčem, co se nás snaží informovat o existenci určitého druhu zkušeností, které se vyskytují mimo kapacity našeho jazyka a tudíž i mimo náš svět. *Traktát* při tomto čtení reprezentuje jakýsi klíč, který nám má pomoci si tento aspekt uvědomit a odemknout metaforickou bránu. Po tomto činu nám však text přestává být užitečný a dle Wittgensteina by bylo nejlepší ho odhodit („He must so to speak throw away the ladder, after he has climbed up on it.“). Tomuto čtení může být i návodný Wittgenstenův stav v době psaní *Traktátu*. V době formování svých tezí se totiž účastnil bojů první světové války. Ve svých soukromých denících popisuje tyto zážitky jako transformativní a několikrát zmiňuje svojí blízkost smrti. Tou dobou mu dělala společnost kniha od Lva Tolstoje *Stručný výklad evangelia*, která už sama o sobě může vybízet k jistému náboženskému či mystickému stavu mysli. Ať je však naše interpretace *Traktátu* jakákoliv, zůstává tato kniha jednou z nejvýznamnějších publikací v oboru filosofie a logiky. *Traktát* také zůstává jedinou prací, kterou Wittgenstein vydal za svého života[4]. [19–21]

Spletitá podoba Traktátu je postavená na myšlence, že cílem logické analýzy jazyka je dospění k elementárním propozicím. Na počátku 30. let však Wittgenstein začal důrazně odmítat systematické přístupy k jazyku a poznání. V této době Wittgenstein také začal nazývat svou ranou práci jakožto dogmatickou. Přičemž toto označení znamenalo pro Wittgensteina jakoukoliv koncepci, která vytvářela „propast“ mezi otázkou a odpovědí. Což znamenalo, že taktéž upustil od rigidních pravidel, které jsou symbolické pro jeho rané období a tím také dogmatismu. Období mezi *Traktátem* a *Zkoumáními*, čili mezi raným a pozdním obdobím, může být chápáno jakožto realizace tohoto odmítnutí a jeho následků. Přechod od idealistické říše logiky do gramatiky každodenního jazyka, od důrazu na analýzu a definice ke konceptům jazykových her a rodinné podobnosti, od systematického psaní k aforismům. A právě tento přechod vrcholí v textu *Filosofická zkoumání*. Po tomto přechody by Wittgensteinovo odpovědí na Sokratovskou otázku po podstatě poznání bylo,

[4] Citace v tomto odstavci pocházejí z prvního anglického překladu C. K. Ogdena

že nemá žádnou podstatu, žádnou esenci. Bylo by tedy chybné si myslet, že by někdo byl schopný podat jedinou systematickou odpověď. (If I was asked what knowledge is, I would list items of knowledge and add ‘and suchlike’. There is no common element to be found in all of them, because there isn’t one). Ve své pozdější filosofii proto tvrdí, že to nejlepší co můžeme v takovémto případě je provést elementární výklad. Uvést příklady použití slova s cílem dokázat, že je nesmyslné chytit se pouze jedné definice. Tedy uvést příklady, v jakém kontextu dává smysl užití slova poznání, a tím tak navrátit slova z jejich metafyzické sféry do sféry každodenního užití.[21–23]

Filosofický pohled, který Wittgenstein následně často kritizuje, se dá označit jako „pohled odnikud“. Jedná se v podstatě o stanovisko, které se prezentuje jakožto platné ve všech situacích a kontextech. Ten je podle něj pouhou filosofickou fikcí. Tvrdí, že tato fikce právě naopak vždy začíná od velmi specifického „někde“ a opírá se o dobře známé příklady, objekty a aktivity. Wittgenstein kritizuje, když se partikulární tvrzení zobecní na platné vždy a všude. Tento vztah života, jazyka a filosofie je pro Wittgensteina centrální skrze celou jeho práci. V tomto ohledu chápe správné užití filosofie jakožto aktivitu určenou k objasnění a pochopení jazyka.[21, 24]

2. Podoba textu Filosofická zkoumání

Pozdní období je tedy pro jisté autory, a koncepci této práce, synonymní s dílem *Filosofická zkoumání,* které bylo, podobně jako zbytek Wittgensteinovo díla, publikováno až po autorově smrti. Finální podoba tohoto spisu je tedy prací správců Wittgensteinovy pozůstalosti v čele s jeho studentkou G. E. M. Anscombe. V díle Wittgenstein podniká kritiku dosavadní filosofické tradice spolu s myšlenkami, které zastával ve svém raném období. Avšak i tady zůstává jeho hlavním zájmem jazyk, jeho užití a jakým způsobem nás jeho užití formuje.

Už na první pohled se text odlišuje od *Traktátu*. Pryč je ono proslulé číslování propozic (někdy označované jako „genetické“), kdy každá propozice značí myšlenkový rozvoj propozice nadřazené. Namísto něj je každý paragraf postavený na rovnou úroveň ostatních a je označený celým číslem. Liší se také jazyk, jakým je dílo psané. Strohý a praktický styl vystřídaly často ironické aforismy ne nepodobné Nietzscheho stylu. Publikace samotná sestává ze dvou textů *Filosofická zkoumání* (paragrafy 1 až 693) a fragmentu *Filosofie psychologie* (paragrafy 1 až 362). První text byl hotový a připravený k publikaci již za Wittgensteinova života. V roce 1938 jej poskytl k tisku univerzitě v

Cambridge, ale během jednoho měsíce se jej rozhodl nepublikovat. Během následujících let na něm provedl několik úprav až do roku 1950, kdy ho ponechal k posmrtnému vydání. Druhý text, zabývající se filosofií psychologie, byl dle rozhovoru mezi Wittgensteinem a editory *Zkoumání* zamýšlený jakožto náhrada paragrafů 491-693. Jelikož ale není jasné, jakým způsobem by měly být implementované do původního textu, nebo jakým způsobem jimi měl být nahrazen, rozhodli se jej editoři začlenit jakožto „druhou část“, fragment. Důvodem však může být také to, jakým způsobem se druhý text vydává novým směrem v diskuzi konceptu aspektového vidění. [21, 23, 25]

Argumenty *Zkoumání* jsou opakovaně vystavěné následujícím způsobem. Nejprve Wittgenstein vyjádří filosofický postoj, proti kterému bude vystupovat. Poté následuje popis velmi konkrétních okolností, za kterých je tento postoj obhajitelný. Na závěr Wittgenstein přednese pozorování, že okolnosti v příkladu jsou velmi omezené a jakmile se dostaneme do situace mimo ně, přestává být zastávání onoho postoje obhajitelné. Stern tuto strukturu argumentu ve svém rozboru *Zkoumání* označuje za Sokratovskou. Stádia argumentu se také velmi často vyskytují v rychlém sledu za sebou. První příklad můžeme vidět už na úplném začátku *Zkoumání*. Wittgenstein v odstavci §1 prezentuje citaci z *Vyznání* sv. Augustina, ve které nám skrze Augustinova slova představuje referenční teorii významu („When grown-ups named some object and at the same time turned towards it, I perceived this, and I grasped that the thing was signified by the sound they uttered“). Ta nám reprezentuje první krok argumentu, neboli představení filosofického postoje. V §2 nás Wittgenstein vybízí, abychom si představili jazyk, pro který by tento princip fungoval. Jako příklad uvede jednoduchou jazykovou hru stavitelů, jejichž řeč se skládá z označení pro druhy kamene. Stavitel A řekne název kamene a stavitel B ho podá. Takto popsaná situace nám vyznačuje velmi specifický příklad fungující teorie z §1. V následujícím §3 Wittgenstein doplňuje, že Augustin sice předvádí systém komunikace, to však neznamená, že je nám v něm poskytnuto vše, co bychom tradičně nazvali jazykem. Můžeme například narazit hned na několik případů, kdy se v situaci popisované §2 musíme zeptat „Je toto označení správné?“. Můžeme tedy vidět, že tato struktura argumentu je nám představená už v prvních momentech naší interakce s textem a Wittgenstein s ní nadále pracuje napříč celým zněním *Zkoumáni*. V určitých pasážích se však stane, že struktura argumentu je narušená a změní svojí podobu. Změny se týkají především umístění třetího kroku argumentu (protinázoru). V několika instancích se tento krok vyskytuje přímo před uvedením kroku dva. Stern tyto výjimky interpretuje tím způsobem, že Wittgenstein nemá

v úmyslu vyřešit filosofické problémy, ale celé je rovnou „rozpustit“ („...not to solve philosophical problems, but to undo or ‘dissolve’ them...“). Řešení je nám poskytnuté ještě před střetnutím s problémem. Wittgenstein nám tak názorně ukazuje, že tu v první řadě vůbec žádný problém nebyl. V jiných případech ale třetí krok argumentu chybí. Je uvedený pouze „výchozí pohled“ bez rozřešení. Stern tak spekuluje, že absence Wittgensteinovy odpovědi má značit cvičení pro čtenáře. Je totiž na něm, aby argument domyslel za sebe.[21, 23, 25]

Wittgenstein často také používá dodatečnou metodu ke třístupňovému argumentu, která přímo vyplývá z jeho užití. Po provedení argumentu, totiž často podrobněji přezkoumává příklad z druhého kroku. Tím nám dle Sterna chce ukázat, jak se na první pohled pevně definované koncepce samy rozplétají pod trochou tlaku. Ani ty nejpřímočařejší a nejočividnější věci se nám po chvíli zkoumání mohou jevit jako nesmyslné. Pokud se přidržíme příkladu z §1, může se nám zdát naprosto přirozené, že slova zastupují předměty. Dodatečným výkladem se nám však Wittgenstein snaží tyto na první pohled očividné věci odcizit a tím nám ukázat, jak moc jsme je brali jakožto samozřejmé. Přestože Wittgenstein říká, že bychom se mohli pokusit představit si, že příklad v §2 konstituuje celý jazyk stavitelů A a B, z následujících paragrafů vyplývá, že není zcela jasné, zda je to možné. Můžeme si představovat scénáře, ve kterých se odehrávají interakce oněch stavitelů. Jen stěží bychom ale mohli pochopit společenství, které by mělo pouze lingvistické schopnosti popsané v §2.[21]

Navzdory tomu, že *Zkoumání* mají podobu souvislého textu, který není dělený do kapitol, ale pouze do číslovaných paragrafů (v případě první části), můžeme v nich zpozorovat clustery myšlenek, či konceptů. Jedná se vždy o několik paragrafů, které se vždy vyjadřují k jedné myšlence. Každý shluk má vždy jasně vyznačené téma, které se v něm bude rozvíjet. Můžeme je proto chápat jako určité kapitoly, které nám můžou pomoci s orientací v textu. Mezi těmito shluky také existují explicitní odkazy na dřívější argumenty, které tato témata spojují dohromady. Je tak možné sledovat linii návaznosti napříč celým textem, přičemž téměř všechny nějakým způsobem vedou zpět k §2. Jednotlivé kapitoly nejsou však ostře oddělené a v některých případech se témata prolínají.[21, 23, 25]

Stern ve svém úvodu do *Filosofických zkoumání* vyznačuje pět hlavních částí (kapitol), z nichž první tři se vyjadřují o jazyku a zbylé o filosofii mysli. První kapitola (§1–64) je kritikou Augustinovy koncepce jazyka, dle které slova získávají svůj význam označováním věcí (jakkoliv je to nepřesné vůči Augustinovi). Skrze tuto kapitolu se Wittgenstein také

vymezuje proti teorii významu prozkoumané v *Traktátu*. Po této kapitole by čtenáře mohlo napadnout, že jestliže význam nepramení z objektů mimo jazyk, pak musí pramenit z jazyka samotného. Z toho důvodu druhá a třetí kapitola (§65–242) kritizují názor, že lingvistická a logická pravidla jsou základem pro význam slov. Čtvrtá kapitola (§243–427) začíná kritikou konceptu soukromého jazyka a související myšlenky, že naše psychologické koncepty získávají svůj význam odkazováním na objekty v naší mysli. Mezi diskutované psychologické koncepty patří: pocit a vizuální vjemy, myšlení, představivost, „já“ a vědomí. Pátá a poslední kapitola (§428–693) je zaměřená na více menších témat: intencionalita, negace, porozumění, zamýšlení a chtění. Je však samozřejmě možné tyto zmíněné části rozdělit ještě jemnějším stylem, podle jednotlivých konceptů. Pro ilustraci nám může posloužit alternativní dělení, které provádí Glock ve *Wittgenstein's Dictionary*.[21, 23, 25]

Při čtení textu *Zkoumání* můžeme pozorovat, že Wittgenstein často přerušuje proud svého výkladu. Ať už vložením dodatečného vysvětlení, či uvedením protinázoru. Za tímto účelem často používá symbolu dlouhé pomlčky „—“ (či její kombinace se standardní pomlčkou „——“). Stern a jiní autoři tyto vložené promluvy označují jakožto jednotlivé hlasy či jiné perspektivy. V textu tak může vznikat zdání dialogu mezi dvěma či více hlasy. Tradičně je hlavní perspektiva, která pohání výklad kupředu, označovaná jako „vypravěč“ (narrator) a mnohdy také bývá určitými autory (například Jane Heal) ztotožňována s Wittgensteinem. Wittgensteinovi jsou tak připisovány názory, které zastává vypravěč. Stern však uvádí, že Wittgensteinovy intence nejsou v tomhle ohledu zcela jisté. Z tohoto důvodu ve svém uvedení do *Zkoumání* připisuje tento hlavní argumentační proud „Wittgensteinovu vypravěči“, než Wittgensteinovi samotnému. Další hlas, který v textu identifikuje je „Wittgensteinův tazatel“ (interlocutor). Opět si můžeme všimnout, že se jedná o „Wittgensteinova tazatele“ a nikoliv o „tazatele“. Stern tak stejně jako v předchozím případu dává najevo, že není jasné, které názory Wittgenstein opravdu zastával a které jsou přítomné pouze pro možnost plné argumentace. Nechce však říct, že si nemůžeme být jistí tím, co si Wittgenstein myslel. Pouze naznačuje, že se text nedá číst úplně černobíle. Musíme však brát ohled i na to, že identifikace a i označení těchto hlasů je až výsledkem interpretace textu. V díle samotném nejsou nijakým způsobem označené, a proto si nemůžeme být jistí, kdy jaký hlas promlouvá ani dokonce kolik jednotlivých hlasů se v díle vyskytuje. Stern proto navrhuje, aby se dialog chápal jakožto výměna mezi neurčitým počtem hlasů, přičemž žádný z nich nemůžeme zcela ztotožnit s názory

Wittgensteina. Alespoň ne bez problémů. Hlas, který posouvá diskuzi, tak Stern označuje za „Wittgensteinova vypravěče“ a hlas který vstupuje do výkladu a vznáší námitky za „Wittgensteinovo tazatele“. Dle Sterna neukazuje autor *Zkoumání* své pravé názory v podobě Vypravěče, ani v podobě Tazatele. Naopak dle něj můžeme tyto názory zahlédnout v momenty, kdy ironické a místy i agresivní hlasy ustoupí a naše pozornost je navedena na pozoruhodné přirovnání, nebo na problematiku, kterou ostatní filosofové neberou dostatečně vážně. Protože právě to je dle Sterna Wittgensteinův hlavní záměr. Ne skrze své dva prostředníky argumentovat pro a proti, nýbrž nám ukázat něco o smyslu debaty samotné. Pokud bychom se zaměřovali na dialog mezi hlasy, mohli bychom jako hlavní Wittgensteinův přínos chápat rozlišení mezi bezproblémovým každodenním jazykem a jazykem metafyziky, který zapříčiňuje veškeré problémy. To však dle Sterna není Wittgensteinovým zájmem. Wittgenstein se nám nesnaží říci, že řešení skeptického paradoxu či soukromý jazyk jsou a nebo nejsou možné, ale naopak, že debata samotná nenese žádnou hodnotu. Jeho diskuze filosofických problémů nemá být podpora jednoho z uváděných názorů, nýbrž jejich vzájemná negace, která nás zanechá s prázdnýma rukama.[21, 23]

V rámci této práce jsem pracoval se čtvrtou, kritickou edici *Filosofických zkoumání*. Tato edice obsahuje jak německý originál textu, tak také anglický překlad provedený G. E. M. Anscombe. Oproti předchozím edicím, bylo však do tohoto překladu zakomponováno několik změn ze strany P. M. S. Hacker a Joachim Schulte. Tyto dvě verze textu jsou v publikaci uspořádané takzvaně *en face*, kdy nejprve se objevuje německý originál a následně jeho anglický překlad. Tato edice obsahuje dvě části. První je text *Philosophische Untersuchungen* (*Filosofická zkoumání*) tak, jak je napsal a zeditoval Wittgenstein. Jednotlivé promluvy v této části jsou vždy označené „§“ a arabskou číslicí (od 1 do 693). Druhá část je pak nazvaná *Philosophie der Psychologie — Ein Fragment* (Filosofie Psychologie — Fragment). Tato část byla v dřívějších edicích známá opravdu jako „*Část II*“. Tato část je složena z Wittgensteinových poznámek, které byly nejspíše zamýšlené jakožto náhrada určitých pasáží z části filosofických zkoumání. Text této části je taktéž číslovaný za pomoci arabských číslic, což je novinkou této edice. Určité segmenty jsou však také ohraničené římským číslováním, které bylo přítomné i v dřívějších edicích. Symbolizují tak, jednotlivé fragmenty textu. Novinkou této edice jsou taktéž podrobné editorské poznámku vysvětlující rozhodnutí překladatelů, spolu s některými odkazy a aluzemi vyskytující se ve Wittgensteinově originálním textu. Další

novou sekcí v této edici, je také esej popisující historii textu *Filosofická zkoumání* a nesnáze pojící se s překladem Wittgensteinových textů.[23]

3. Koncepty Filosofických zkoumání

Leitmotivem, který se proplétá skrze většinu *Zkoumání,* je především jazyk a jazyková reprezentace. Jak už bylo řečeno, toto téma Wittgensteina zajímalo již při práci na *Traktátu*. Tématem je zde však upuštění od dogmatického přístupu a navržení nové alternativy, která je inspirovaná naším denním užitím jazyka. [25]

3.1 Kritika Augustinovy koncepce jazyka a vymezení proti Traktátu

Text *Filosofická zkoumání* začíná citátem od sv. Augustýna, který původně pochází z textu *Vyznání*. V citátu Augustýn popisuje, jak se jako dítě učil jazyk. Stěžejní částí v citované pasáži je: „Every word has a meaning. This meaning is correlated with the word. It is the object for which the word stands“[5]. Důležité je zmínit, že Wittgenstein nevnímal Augustýnův názor jakožto ucelenou teorii jazyka, nýbrž jako proto-teoretické paradigma. Navzdory tomu si Wittgenstein myslel, že si pasáž zaslouží kritickou pozornost, jelikož zmiňuje principy, které leží v základech sofistikovanějších teorií významu. Glock ve svém *Slovníku* identifikuje následující myšlenky z §1, jako důležité: „každé individuální slovo má svůj smysl“, „všechna slova jsou jména a zastupují objekty“, „významem slova je objekt, který reprezentuje“, „spojení mezi slovy a významy je ustanoveno ostenzivní definicí“ (Přestože Wittgenstein v §6 uvádí, že tento jev nechce nazývat ostenzivní definicí, protože dítě se ještě nemůže ptát na jména.), „věty jsou kombinacemi jmen“. Z nichž jsou následně vyvozené tyto důsledky v pozdějších paragrafech: „Jedinou funkcí jazyka je reprezentovat realitu. Slova referují a věty popisují.“ „Dítě dokáže vytvořit asociaci mezi světem a objekty čistě skrze myšlení, což znamená, že musí ovládat soukromý jazyk, aby bylo schopné se naučit jazyk veřejný“. Wittgenstein nám tak skrze Augustýna představuje referenční teorii významu v níž každé slovo získává svůj význam skrze reprezentaci objektu ve světě. Sdílí tak velikou podobnost s koncepcí jazyka, kterou Wittgenstein zastával během doby, kdy pracoval na textu *Traktátu*. Na toto pojetí odpovídá Wittgensteinův vypravěč, že akt pojmenování sám o sobě nic neznamená, jelikož veliká část této interakce závisí na situaci která předcházela a co následovala. Stěžejní myšlenkou je zde tedy to, že různé způsoby použití slova, závisí na kontextu ve kterém se nacházíme. [21, 23, 25]

---

[5] Překlad uvedený v textu *Filosofických zkoumání* v poznámce pod čarou.

3.2 Řečové hry

Jak už bylo řečeno, pojetí jazyka přítomné ve *Zkoumáních* je velice odlišné od Wittgensteinových dřívějších představ o fungování jazyka. V §81 tvrdí, že během jednoho rozhovoru s F. P. Ramseyem diskutovali podobnost formální logiky a normativní vědy. Tato podobnost plyne z faktu, že logika operuje se slovy za použití kalkulu s fixními pravidly. Wittgenstein dále zmiňuje, že tato podobnost dále implikuje podobnost s hrami, které také mají určitá fixní pravidla zajišťující jejich hladký průběh. V textu *Zkoumání* je koncept jazyka jakožto řečové hry představený několikrát za pomoci příkladů (jak skutečných, tak imaginárních). Prvním takovým příkladem může být už imaginární společnost stavitelů z odstavce §2, hra „ring-a-ring-a-roses“ odstavce §7, či dokonce jak se děti učí slova v citaci Augustina (ptaní-se a dávání odpovědí). Wittgenstein také v §7 označuje celý jazyk (míněno jako koncept) a všechny ostatní aktivity, kterých je jazyk součástí, za řečovou hru. Nemá zde však na mysli pouze akt vyslovování slov, ale i vše ostatní, co se k tomu pojí. Obklopující okolnosti jako: vzorce akcí objevující se během vyslovování slov, předměty v našem okolí, místa, na kterých se slova pronášejí, a naši „spoluhráči“ těchto jazykových her (obecně je tyto okolnosti možné shrnou jako kontext). Všechny tyto okolnosti mají vliv na pravidla a průběh jazykové hry, kterou můžeme hrát (ale také nemusíme). Nicméně, Wittgenstein nechápe řečové hry jakožto základ pro nový model jazyka. Jeho úmyslem není nahradit stávající model logicko-matematického kalkulu, ale spíše nabídnout možnost srovnání pro složité případy. Poskytnou nový způsob, jak nahlížet situaci, ve které se nacházíme, abychom si snáze ujasnili podobnosti a rozdíly mezi jednotlivými jevy (§130). Wittgenstein nám tedy nechce říci, že jazyk není nic než řečová hra, nebo že můžeme pravidla svého jazyka měnit tak snadno jako pravidla her. Spíše pouze poukazuje na paralelu mezi jazykem a hrami, která nám může pomoci pochopit aspekty použití našeho jazyka. Aspekty, které jsou často nepostřehnutelné v rámci jiných přístupů. Jeho cílem je totiž umožnit nám nahlížet tyto problémy z nového úhlu, který by nám dle Wittgensteina měl pomoci rozptýlit nejasnosti, které vyvstávají za použití tradičních přístupů k jazyku. Zmiňuje, že vina v tomto ohledu většinou spadá na rigidní koncepce jazyka, které až přespříliš spoléhají na abstraktní systémy pravidel. Za jejich použití tak dle Wittgensteina trávíme čas především vymýšlením, jak musejí (nutně) fungovat (§51).[21, 23, 25]

3.3 Rodinná podobnost

První nástin rodinné podobnosti uvádí Wittgenstein v odstavci §65. Otázka vznesená v tomto paragrafu je: „jak definovat vlastnost řečové hry, aby tato definice platila pro všechny ostatní příklady řečových her“, či jinak co „můžeme označit za esenciální vlastností řečových her“. Načež vzápětí odpovídá, že tato snaha by byla marná z toho důvodu, že zde neexistuje jediná věc, která by byla společnou pro všechny druhy řečových her. Žádný znak, který by byl společný jím všem. Namísto toho jsou však mezi nimi různé druhy příbuzností. Právě na základě těchto příbuzností, je všechny můžeme označovat jakožto jazyk. V následujícím odstavci §66 použije Wittgenstein příkladu s definicí samotného slova hra. Nyní ne ve smyslu „řečová hra“, nýbrž tradiční koncepce hry (desková, karetní, pohybová). Vybízí nás, abychom si zkusili představit, jeden znak, který by byl společný pro všechny instance slova hra. Nejprve nás zavede k deskovým hrám se všemi jejich znaky, které se k nim pojí a následně se je pokusí aplikovat na hry karetní. Mezi těmito skupinami se dá najít mnoho společných vlastností, přestože ne všechny se absolutně překrývají. Kupříkladu že je většinou hrajeme ve vnitřním prostředí, nevyžadují moc tělesné aktivity, potřebujeme k nim další hráče. Co ale když do této skupiny přidáme míčové hry? Dle Wittgensteina je očividné, že není moc společných znaků mezi prší, šachy a fotbalem. Můžeme namítnout, že musí existovat vítěz a poražený. Co když ale osamělé dítě hraje míčovou hru, spočívající v házení míče o stěnu. V ten moment nám zmizí i tento znak. Dále je možné trvat na tom, že hry jsou zábavné. Stačí nám však se na chvíli zamyslet, abychom si vzpomněli, že ne každá hra nás nutně bavila, nebo nám její pravidla přišla zábavná. Můžeme tedy zmínit nepřeberné množství znaků, ale je stěží dokážeme uvést jeden, který by byl platný pro všechna použití slova hra. Namísto toho je možné (a možná i zapotřebí) vytvořit sít jednotlivých her a pozorovat, jak tvoří clustery na základě jim společných znaků. V tomto smyslu Wittgenstein míní rodinnou podobnost. Instance jednotlivých konceptů jsou pro něj jako členové rodiny určitého konceptu. Mezi jednotlivými členy jsou určité podobnosti a určité rozdíly (výška, postava, barva očí a barva vlasů, povaha), které se míchají a překrývají v nejrůznějších kombinacích.

Wittgenstein dále v odstavci §67 rozebírá: „proč můžeme něco nazvat číslem?“. Načež poskytuje odpověď, že je to kvůli přímé příbuznosti s něčím, co jsme až doposud nazývali číslem. Tím tato věc získává nepřímou příbuznost s dalšími věcmi nazývanými čísla. Tímto způsobem se tak rozšiřuje pole působnosti našich konceptů. Tento přístup však není jediný, skrze který můžeme přistupovat k vymezení pojmů. V následujícím odstavci §68

uvádí, že můžeme označení „číslo“ použít pro pevně vymezený pojem, ale zároveň můžeme to stejné označení použít i pro pojem, který není vymezený. Tak ostatně dle Wittgensteina používáme slovo „hra“. Pokud po nás totiž někdo vyžaduje definici hry, tak jsme schopni si nějakou vymyslet. Popřípadě, dokážeme uvést několik příkladu a říci: „Toto jsou hry“. [21, 23]

3.4 Paradox následování pravidla

Wittgenstein nám v paragrafu §185 nastíní následující případ. Vybízí nás, abychom si představili žáka, který dostane za úkol pokračovat v sérii čísel. Je mu dáno číslo 1000 a instrukce, aby k němu přičítal vždy „+2“. Žák tedy poslechne a následně pokračuje v sérii: 1004, 1008, 1012. Žák je za svojí chybu pokárán, ale nechápe, že postupoval chybně. V jeho očích je tato série čísel správně. Wittgenstein se skrze tuto část *Zkoumání* zamýšlí nad konceptem následování pravidla a myšlenku že „pravidlo může být interpretováno různými způsoby v každém případě jeho užití“. Pokud se setkáme s určitým matematickým pravidlem (například to z §185) či jazykovou konvencí, jak poznáme, že se jím lidé řídí správně? Abychom jednali správně podle pravidel z paragrafu §185, musíme vědět, co znamená „2“, „přičítat“ a jiné koncepty. Toto pravidlo je tak složené z dalších pravidel. Jak si tedy můžeme být jistí, že i tato jsou aplikovaná správně? To následně vede k nekonečnému regresu, při kterém zjišťujeme, že si nemůžeme být jistí, zda je jakýkoliv koncept či pravidlo aplikované správně. Z toho pro Wittgensteina v paragrafu §201 plyne následující, a to že: „žádný způsob jednání nemůže být určený pravidlem, protože každý způsob jednání může být uveden do souladu s pravidlem“ a také: „jestliže lze každý způsob jednání uvést do souladu s pravidlem, pak jej lze také uvést do rozporu s ním“. Pokud budeme tato tvrzení brát jakožto premisy, dostaneme se k paradoxu. Wittgensteinovou odpovědí na tento paradox je opuštění od rigidních pravidel směrem k něčemu více fluidnímu. Následování pravidla by tak mělo spíš být chápáno jako zvyk (praxe) na základě kontextu a použití v rámci určité jazykové hry, kterou hrajeme.[21, 23]

3.5 Paradox soukromého jazyka

V paragrafech §243–268 Wittgenstein diskutuje myšlenku soukromého jazyka. V principu se jedná o jazyk, kterému rozumí pouze jeden jediný člověk. Tento jazyk tedy slouží pouze k popisu toho, čím jsou bezprostřední soukromé pocity a myšlenky onoho člověka. V paragrafu §261 Wittgensteinův vypravěč argumentuje, že slovo potřebuje pro své užití odůvodnění, kterému mohou všichni rozumět. Myšlenka soukromého jazyka, kterému

rozumí pouze jeden člověk je tedy dle něj už ze své podstaty nemožná. Tvrdí, že jazyk je v zásadě veřejnou činností, která si zakládá na sdílených konvencích a významech v rámci společnosti. Znaky jazyka mohou fungovat pouze v momentu, kdy zde existuje možnost posouzení správnosti jejich užití a případná korekce jejich užití. Namítá tak, že jazyk popisující čistě subjektivní zážitky by mohl znamenat popření základních funkcí jazyka. Wittgenstein se ale nezastaví pouze u této kritiky. Dále totiž prozkoumává myšlenku, zda by bylo vůbec možné takovýto jazyk vytvořit (nehledě na porušení principů jazyka jako takového). Přesněji, jakým způsobem jsou slova tohoto jazyka napojené na jeho vnitřní objekty (jelikož tento jazyk popisuje bezprostřední a soukromé pocity). Wittgenstein nejprve zdůrazní, že běžné pocity či zkušenosti (jako pocit bolesti) nemůžou být předměty soukromého jazyka. To z toho důvodu, že slova označující tyto druhy pocitů a zkušeností jsou už součástí našeho veřejného jazyka. Wittgenstein dále argumentuje, že i zdánlivě soukromé zážitky jsou často zakořeněné ve sdílených zvycích. V tomto smyslu tedy nemůže být soukromý jazyk, soukromý v silném smyslu, který by si přáli jeho proponenti. Stern poznamenává, že v paragrafu §256 Wittgenstein znovu aplikuje argument z paragrafů §2 a §3. Čili uvádí, že soukromý jazyk by měl, stejně jako řečová hra společenství stavitelů, velice omezenou působnost. To by proponentům tohoto druhu jazyka v zásadě nevadilo (omezená působnost vyplývá už z konceptu soukromého jazyka), tato poznámka nás však zavede k závažnějšímu nedostatku. Podle Wittgensteina hlavní problém existence soukromého jazyka věží v jeho nedostatku aplikovatelnosti a koherence. Jelikož je soukromý jazyk koncipován jakožto autonomní systém reprezentace, kterému rozumí pouze jeho mluvčí, dostávají se jeho proponenti do problému, který se nazývá „*Paradox soukromé ostenze*“. Ten v podstatě znamená, že jelikož má tento jazyk pouze jediného uživatele, může být ostenzivní definice rozdílná v každém případě. Ve výsledku je tedy nemožné, aby znak v soukromém jazyce cokoliv kdy znamenal, jelikož mu nikdy nebyl dán význam. Způsob užívání slova je totiž praxí, jazykově strukturovaným postupem. Jediným způsobem, jakým užití slova může být soukromé, je ve smyslu, že se jej rozhodnu držet v tajnosti. Chápat jej ale jako „nutně soukromé“ znamená špatně chápat naše užívání jazyka. Pouhé „myšlení si“, že člověk dodržuje pravidlo, nestačí k tomu, aby bylo možné tvrdit, že jej opravdu dodržuje. Wittgenstein na závěr uvádí, že proponenti soukromého jazyka jsou nuceni zvolit jednu z následujících dvou alternativ (obou pro ně neuspokojivých). První možností je přijmout, že se jedná o pouhý koncept soukromí v tradičním slova smyslu a se kterým běžně operujeme (držet způsob užití v tajnosti). Tato koncepce by znamenala, že s ním není možné operovat jakožto s filosofickým konceptem

tak, jak by si to přáli jeho proponenti. Druhou možností je přiznání, že se jedná o „filosofický superkoncept“ vytvořený na míru za účelem podpoření vlastní teorie, který je ale ve velké míře odtržený od reality užití jazyka.[21, 23]

1. Smysl modelování přirozeného jazyka

Můžeme si klást otázku „co nám tyto pravděpodobnosti vypovídají o jazyku“. Může nám takováto statistika analýza textu poskytnout nějaké relevantní informace? Je zřejmé, že některé poznatky plynoucí z těchto pravděpodobnostních analýz nebudou překvapivé. Například, že za přídavným jménem se bude nejspíše vyskytovat podstatné jméno. Tato zjištění můžeme zařadit spíše do lingvistické kategorie. Z modelu však mohou plynout také jisté kulturní poznatky, například že zájem o „čínskou kuchyni“[6] bude spíše v ne-čínských zemích. Dále se mohou ukazovat různé kulturní zvyklosti, jako je způsob, jakým začínáme rozhovor nebo jakým ho ukončujeme. Co je však pro obor filosofie zajímavější, je možnost vytvořit model specifického korpusu. Ten může představovat jednotlivá kniha, celé dílo určitého autora, všechny publikace pramenící z filosofického směru (francouzský existencialismus 20. století), filosofického uskupení (Bádenské novokantovství), či dokonce souhrn textů vyjadřující se ke stěžejnímu konceptu (jsoucno). Tento na první pohled „suchý“, či matematický přístup tak může poskytnout nový způsob analýzy a interpretace filosofických textů. Akademik se nemusí „omezovat“[7] na tradiční komparativní metodu či pouhou interpretaci textu. Zde se mu naskytuje možnost rozšířit svůj arzenál o víceméně „objektivní“ nástroj pro analýzu textu. Samozřejmě, že analýza jazykové stránky díla byla možná i před příchodem jazykových modelů. Žádala si však mnohonásobně větší množství času, než za využití metod strojového učení. Akademik si tak může nechat jednoduše analyzovat text a pak následně zasvětit více času interpretaci vzniklých dat. Tato metoda nám totiž může ukázat zajímavé okolnosti, v rámci kterých autoři píší. Je pro nás tak možné odkrýt jen málokdy zřetelné struktury, které tvoří základ filosofického díla. Je však dobré mít stále na paměti, že pravděpodobnostní rozdělení vyplývá pouze z daného korpusu, který jsme analyzovali. To znamená, že závěry, které bychom z pravděpodobností vyvozovali, budou platné pouze pro daný korpus, který jsme analyzovali. Za příklad nám může sloužit článek z roku 2017, *A Visual Representation of Wittgenstein’s Tractatus Logico-Philosophicus*. Ten byl publikovaný skupinou komputačních lingvistů (computational linguists), která si dali za cíl porovnat podobnost číslovaných propozic, napříč jednotlivými překlady Wittgensteinova *Traktátu*. Jejich metodou byl především výpočet kosinové podobnosti, mezi tokeny jednotlivých propozic.

[6] Je pravděpodobné, že čínsky psaná literatura, nebude k čínskému jídlu odkazovat jakožto „čínskému jídlu“, ale jednoduše jakožto k „jídlu“.

[7] Není myšleno pejorativně.

Už však z této poměrně konceptuálně jednoduché metody, je možné extrahovat zajímavé informace. Tato analýza totiž ukázala, že je zde silnější topologická podobnost mezi německým originálem a prvním anglickým překladem (Ogden a Ramsey), než překladem z roku 1961 (Pears a McGuiness). Tyto zmíněné metody pro práci s textem jsou již dobře známé v oboru NLP, ačkoliv v oboru filosofie se jejich užití stále nerozšířilo (mimo směr digital humanities).[7, 26]

Další příkladem, k čemu se jazykové modelování dá využít, jsou generativní modely. Toto téma již bylo nastíněné v sekci A. Jedná se systémy (v dnešní době se většinou tvoří za využití transformerové architektury), které generují text na základě pravděpodobnosti výskytu dalšího znaku. Jedná se většinou o velké jazykové modelu, tvořené na velkých datasetech. To především z důvodu, že tyto modely se používají ke generování textu, jehož hlavním znakem má být nerozpoznatelnost od lidmi tvořeného textu. Jsou zde tedy uváděné v praxi myšlenky: „čím větší dataset, tím sofistikovanější bude modelem generovaný text“ a „čím diverznější dataset, tím bude model schopný generovat více všestranný text“. Tyto modely je pak následně možné za pomoci menších datasetů dotrénovat na specifické úlohy. Jedním takovým úkolem může být také převzetí určitého literárního stylu. V České republice na podzim roku 2019 probíhal projekt *Digitální filosof*, který si dal za úkol vytvoření hned několika jazykových modelů, které by převzaly jak literární styl, tak stěžejní filosofické myšlenky daných filosofů. Ke zpracování tohoto projektu byl využitý model GPT-2 (345M parametrů) a filosofické texty několika autorů (Hannah Arendt, Gilles Deleuze & Félix Guattari, Michel Foucalt, Václav Havel, Tomáš Sedláček, Peter Singer). V první polovině roku 2023 dále proběhl konceptuálně podobný pokus za využití novějšího modelu GPT-3 (175B parametrů) a filosofického díla Daniela Dennetta. Toto zpracování má oproti zmíněnému českému projektu dvě nezměrné výhody. První je sofistikovanost jazykového modelu použitého ke trénování (GPT-3 dosahuje až 500x většího počtu parametrů). Druhou výhodou je spolupráce filosofa na projektu. Nejenže Dennett poskytl autorům kopii celého svého digitálního díla (15 knih a 269 článků), ale podílel se i na vyhodnocení výsledku projektu. V závěrečné fázi vyhodnocení tak poskutoval komentáře ke generovanému textu. Byla také provedena určitá obdoba Turingova testu, kdy odborníci na Dennettovo dílo a reposndenti blogu (The Splintered Mind) měli za úkol rozpoznat, které citáty jsou generované a které originální. Přičemž výsledkem experimentu byly následující statistiky „Blog readers showed a pattern of mistakes similar to that of the experts, with the highest percentage of correct answers on

the Chalmers and Fodor questions (84% and 52%, respectively) and the lowest percentage on the Robot and Free Will questions (both 35%)“. Na základě těchto poznatků můžeme tedy předpokládat, že je možné tímto způsobem dosáhnout natrénovaného jazykového modelu, který je schopný generovat text, jenž je těžké rozeznat od „originálu“. A to i pro experty v daném oboru a na dané téma.[10, 18, 27]

Je však důležité zmínit, že tyto myšlenky nejsou všeobecně přijímané a k problematice těchto obrovských datasetů, velkých jazykových modelů a smyslu generovaného textu, se negativně vyjadřuje konferenční příspěvek „*On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜“. Jejich kritika pramení z několika různých úhlů, přičemž prvním bodem jsou ekologické dopady velkých jazykových modelů. V moment vydání nového jazykové modelu je často první informací, kterou o něm slyšíme, jeho parametrická velikost. Ač se nám tento druh informace může často zdát jako marketingový tah (ve smyslu „nový a větší model“), bylo ukázáno, že s velikostí parametrů často roste sofistikovanost jazykového modelu (to však nemusí platit nutně). Autoři tohoto příspěvku si však kladou otázku, zda benefity plynoucí z větší parametrizace modelu, převažují potenciální environmentální rizika plynoucí z energetické náročnosti trénování modelů o této velikosti. Jedním z možných způsobů, jak dohlížet na tyto následky je dle autorů, vždy uvádět energetickou náročnost trénování. Výsledky výzkumu by pak měly být dávány do kontrastu s množstvím konzumovaných zdrojů. Dalším bodem kritiky je, že s velikostí jazykových modelů roste i obtížnost jejich vysvětlitelnosti (explainability). A to nejen z hlediska jejich rozhodovacího procesu, ale také z hlediska zastoupení biasů v trénovacím datasetu. Čím větší máme dataset, tím větší je šance výskytu hegemonních názoru, které mohou negativně ovlivnit marginalizované skupiny. Z větší velikosti datasetu také plyne horší detekovatelnost takovýchto biasů. Co víc, u moderních velkých jazykových modelů, založených na transformerové architektuře není rozhodovací proces transparentní. Nemáme tak jistotu, na základě jakých vypozorovaných vzorců se systém rozhoduje. Finálním bodem kritiky ze strany autorů jsou pak limity velkých jazykových modelů. Tyto modely jsou totiž veřejnosti často prezentované mimo kontext (Pro příklad se nám stačí podívat na webové stránky OpenAI a jakým způsobem prezentují své nejnovější modely). Autoři tohoto příspěvku tvrdí, že jazykové modely neprovádějí „natural language understanding“ (pochopení přirozeného jazyka) nýbrž jenom manipulaci s lingvistickou formou. Na základě všech těchto nedostatků autoři provádějí kritickou skepsi vůči přínosu

neustále rostoucí velikosti jazykových modelů. Zcela přitom však tuto technologii nezavrhují, pouze poukazují na nutné vytyčení výzkumných cílů.[11, 12, 28]

2. Praktické zpracování

V této části práce budu popisovat svůj postup (a jeho ospravedlnění), při fine-tunování velkého jazykového modelu GPT-NeoX za pomoci textu Ludwiga Wittgensteina *Filosofická zkoumání*. Podnětem pro tuto diplomovou práci byl především projekt Dity Malečkové a Jana Tyla *Digitální filosof*. Metodologicky se také přidržuji publikace *Creating a large language model of a philosopher*, ve které se autoři pokoušeli o fine-tuning velkého jazykového modelu GPT-3 na celém díle Daniela Dennetta.[18, 27]

Motivací pro vytvoření jazykového modelu Wittgensteinovy pozdní filosofie je hned několik. Jednak jej můžeme jednoduše označit za jednoho z nejvýznamnějších filosofů 20. století. Jeho přínosy k oborům logiky, filosofie jazyka a filosofie mysli jsou nepopiratelné. Jakožto zajímavým aspektem vytvoření tohoto jazykového modelu, je tak jistá cykličnost. Wittgenstein položil základy, pro stěžejní myšlenky, které ovlivňují obor NLP dodnes. Myšlenky jako „významem slova je způsob jeho použití“, či celý Wittgensteinův koncept rodinné podobnosti. Tato práce zaměřená na vytvoření jazykového modelu, by také mohla sloužit, jakožto podklad pro budoucí výzkum, jehož cílem by mohlo být modelování celého Wittgensteinova díla. To by mohlo být zajímavé již v ohledu, jak fragmentované Wittgensteinovo dílo je. Jak jsem již uváděl v předchozí části, Wittgenstein za svého života publikoval pouze jediné dílo, a to *Traktát logicko-filosofický*. Zbytek setrvává v podobě poznámek, z nichž některé byly zeditované a vydaté správci jeho pozůstalosti. Mohl by být proto zajímavý podnik, vytvořit jazykový model kompletního Wittgensteinova díla.

2.1 Popis použitého modelu

GPT-NeoX je autoregresivní velký jazykový model vyvinutý non-profit skupinou EleutherAI. Tento model má 20 miliard parametrů z čehož 19,9 miliard jsou „ne-embeddingové“ parametry, 44 vrstev, hidden dimension o velikosti 6144 (v podstatě počet hodnot obsažených ve vektoru) a 64 hlavic. Jeho architektura byla z velké míry převzatá z modelu GPT-3 vyvíjeného společností OpenAI. GPT-NeoX používá rotary positional embeddings (rotační poziční embeddingy) namísto learned positional embeddings (naučené poziční embeddingy) modelu GPT-3. Rozhodli se tak na základě dřívějších kladných zkušeností při trénování velkých jazykových modelů za použití této metody. Její implementace má za následek pokřivení embeddingového prostoru. To takovým způsobem, že attention tokenu

na pozici *m* k toneku na pozici *n* je lineárně závislá na $m - n$. Přestože někteří výzkumníci aplikují tyto embeddingy na každý embeddingový vektor, v tomto případě se autoři rozhodli je použít pouze na prvních 25% dimensí embeddingového vektoru. Jejich prvotní experimenty totiž ukazovaly, že toto nastavení má za následek nejlepší vyvážení výpočetního výkonu a efektivity. Další změnou je paralelizace výpočtu attention a feed-forward vrstev a jejich následný součet. Ta byla také motivována pozitivním dopadem na efektivitu. Oproti GPT-3, které ve své architektuře střídá husté a řídké vrstvy, se v případě GPT-NeoX autoři rozhodli použít čistě husté vrstvy, aby se snížila komplexita implementace. Během trénování jazykového modelu se autoři GPT-Neox rozhodli inspirovat nastavením GPT-3. Jelikož jsou velikosti parametrů obou modelů odlišné, rozhodli se autoři interpolovat learning rate GPT-3 13B a 175B, aby dospěli k hodnotě „0.97E−5“ , kterou se rozhodli použít. Po následném testování, se dále rozhodli použít weight decay parametr 0.01, batch size o velikosti přibližně 3.15 milionů tokenů (neboli 1538 kontextů, každý při velikosti 2048 tokenů), 150 tisíc trénovacích kroků, přičemž learning rate se snižoval v rámci kosinového plánu až na 10% (na konci trénování) své původní hodnoty. Při trénování byl použitý „AdamW“ optimizer, který byl rozšířený o „ZeRO“ optimizer za účelem snížení požadavků na paměť při distribuci stavů optimizeru. Z důvodu velikosti modelu implementovali jeho autoři možnost paralelizace tensorů v kombinaci s pipeline paralelismem. Tím jim bylo umožněné rozdělit model napříč vícero GPU, namísto jedné. Během trénování se tedy rozhodli nastavit paralelizaci tensorů na velikost 2 a pipeline paralelismus na velikost 4. Tímto způsobem se jim podařilo dosáhnout účinnosti 117 teraFLOPS na každé GPU.[6, 10, 29]

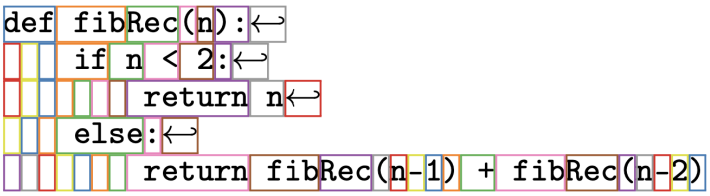
GPT-NeoX byl trénovaný na datasetu the Pile (hromada), který byl taktéž vyvinutý skupinou EleutherAI. Vytvoření tohoto datasetu bylo motivováno čistě pro potřeby trénování velkých jazykových modelů. Trénování tohoto druhu jazykových modelů totiž vyžaduje obrovské množství textu s lidským autorstvím. V době plánování tvorby GPT-NeoX, však nebyl žádný takto veliký dataset jeho autorům volně dostupný. Rozhodli se tak pro tvorbu kompilace volně dostupných internetových dat do jednoho velikého datasetu (během tohoto projektu bylo vytvořeno i několik menších datasetů od třetích stran za účelem jejich přidání do the Pile). Mimo to bylo jejich další motivací demokratizace technologie. Tvůrci datasetu měli totiž obavy z faktu, že přístup k takovému typu dat (ve smyslu velkosti), je mnohdy umožnění pouze velkým korporacím zaměřujícím se na těžbu dat (data mining). Ne z důvodu nějaké arbitrární regulace, nýbrž z důvodu hardwarové a

časové náročnosti shromáždění, anotování a očištění takového velkého objemu dat. Korporace se zaměřením na těžbu dat také svoje datasety téměř nikdy neposkytují zdarma (z očividných důvodů). Dokonce i společnost OpenAI, která se prezentuje svým „open“ (toto přízvisko bylo zejména v posledních letech podrobeno tvrdé kritice ze strany ostatních AI výzkumníků) přístupem k umělé inteligenci, poskytuje pouze datasheet ke svému datasetu, ne dataset samotný. Pro běžného člověka bylo tak do té doby velmi obtížné dostat se k podobnému datasetu. The Pile je tedy složený z několika menších, specializovaných datasetů. Velikost jednotlivých částí (a tedy i procentuální zastoupení v celku) datasetu je dostupná na Github repozitáři datasetu[8]. Pile-CC, je dataset vytvořený za účelem jeho implementace do the Pile a tvoří největší část celkového datasetu. Jeho vytvoření je zásluhou organizace Common Crawl, která se specializuje na „prolézání“ (crawl) dat internetu a tvorbě volně dostupných datasetů a archivů na základě získaných dat. The Pile je tak ve výsledku, dle témat jednotlivých částí, ze kterých je složený, velice diverzním datasetem pro trénování velkých jazykových modelů. Obsahuje nejen obecný text z internetu, ale i články zdravotního výzkumu, beletrii, technické články, repozitáře programového kódu, texty zabývající se právem, patenty, titulky filmů, titulky youtube videí, učebnice aritmetiky, dokumentaci k operačnímu systému Ubuntu a filosofické články. Ve výsledku je tedy dataset the Pile složený z 825 GiB surových textových dat (jeho celková velikost je 1254.20 GiB). Jeho složení reflektuje především záměr jeho tvůrců o vytvoření obecného jazykového modelu. [6, 29, 30]

V případě GPT-NeoX použili jeho tvůrci tokenizér založený na byte-pair enkódingu. se slovníkem o velikosti 50257. Je tak velmi podobný tokenizéru použitém v GPT-2. Liší se však ve třech hlavních ohledech. Za prvé byl tento tokenizér trénovaný na the Pile datasetu a to za účelem konstrukce více obecného tokenizéru (vzhledem k datům na kterých byl trénovaný). Za druhé se tento tokenizér chová jinak k řetězci tokenů neoddělených mezerníkem. A za třetí se tokenizér chová odlišně při výskytu několika znaků pro mezeru za sebou (od 1 až do 24 výskytů včetně). To má za následek, že tokenizér vytvoří menší počet výsledných tokenů. Tyto dvě rozhodnutí nejvíce ovlivňuje proces tokenizace programového kódu (viz. obr.).[6]

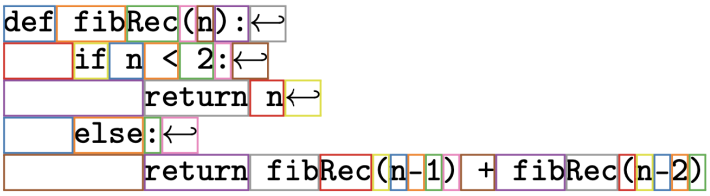
[8] https://github.com/EleutherAI/the-pile

GPT-2



55 tokens

GPT-NeoX-20B



39 tokens

(ukázka tokenizace, obrázek převzatý ze[6])

Tento model jsem si vybral především z ideologického hlediska. Jedná se o open-source model, jehož autoři kladou důraz na demokratizaci AI. Jako takový má navíc uveřejněný dataset na kterém byl trénovaný. Je tedy možné provést případný audit datasetu a prověřit případné biasy, které natrénovaný model může vykazovat. Co víc, model je schopný generovat text, co do sofistikovanosti konkuruje modelu GPT-3. A to i přes několikanásobně menší parametrizaci (20B ku 175B parametrů).

2.2 Výběr a předzpracování trénovacích dat

Jelikož mým cílem je provést fine-tuning velkého jazykového modelu za účelem generování textu, který by připomínal pozdní filosofii Ludwiga Wittgensteina, je výběr mých trénovacích dat očividný. Pro trénování využiji Wittgensteinova díla *Filosofická zkoumání*, konkrétně jeho čtvrtou edici vydanou Wiley-Blackwell v roce 2009. Podoba Wittgensteinova díla byla diskutovaní již v předchozí sekci práce. Z ní vyplývají i mé pohnutky pro volbu trénovacího textu. Přestože, jak už bylo uvedeno, není pochyb o Wittgensteinovo autorství, je zde prostor pro kritiku na základě editorské práce. První část díla *Filosofická zkoumání* je jediným textem Wittgensteinova pozdního období, která byla připravená k publikaci. Jedná se tedy o text, který má takovou podobu, jakou autor zamýšlel. Na základě toho chápu tuto část texty, jako vrcholnou podobu Wittgensteinova pozdního myšlení. Jelikož velký jazykový model GPT-NeoX byl předtrénovaný především

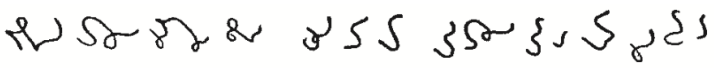
na anglickém textu, rozhodl jsem se pracovat s anglickým překladem obsaženým v tomto vydání.[18, 21, 23]

Předtím, než je možné začít s procesem trénování, je nutné nejprve extrahovat text *Zkoumání* do podoby, ve které s ním je stroj schopný pracovat. Má kopie tohoto textu je ve formátu pdf, což není formát který je pro tento druh operací vhodný. Naštěstí má kopie byla obohacená o kvalitní OCR (optical character recognition) metadata, které není problém extrahovat. To jsem provedl za použití open-source programu *XpdfReader* a jeho funkce *pdftotext*. To mě zanechalo se souborem ve formátu txt, který je pro strojové zpracování vhodnější. Text však po extrakci obsahoval i veškerá dříve skrytá metadata, která nemají s autorovými myšlenkami moc společného. V textu se proto vyskytovala i označení konce stran a jiné „artefakty". Jelikož je tato edice dvojjazyčným vydáním, bylo také nutné odstranit z textu originální německý text. Wittgenstein také v určitých paragrafech uvádí, souběžně s výkladem, citace sv. Augustina, které jsou v latinském originále. Pro latinsky nemluvící čtenáře jsou však přiložené anglické překlady těchto citací. Rozhodl jsem se proto odstranit latinské originály a na jejich místo vložit tyto anglické překlady. Mělo by tak dojít k zachování myšlenkové návaznosti a zároveň umožnění modelu práci s textem. V textu se také vyskytuje několik poznámek editorů, které taktéž byly odstraněny, jelikož se nejedná o Wittgensteinovy myšlenky.

Výsledkem mého zpracování je tedy text *Filosofická zkoumání* ve formátu txt. V této podobě soubor obsahuje všech 693 paragrafů první části textu v anglickém jazyce. Důvody pro nepoužití textu druhé části (fragmentů) byl nastíněný již v sekci B. Především je mé rozhodnutí ovlivněno povahou této části. Není zcela jisté do jaké míry a jakým stylem jí Wittgenstein plánoval zakomponovat do první části textu. Z tohoto důvodu jsem se rozhodl použít jen část první. V mém trénovacím textu také chybí předmluva a poznámky pod čarou. Mým cílem je jednoduše natrénovat model na čisté Wittgensteinově filosofii postihnuté v těchto 693 paragrafech.

Text *Zkoumání* však také obsahuje nemálo obrázků, nebo strojově nezapsatelných symbolů, na kterých Wittgenstein často popisuje své koncepty. Tyto obrázky jsou v texty vždy dostatečně popsané před jejich výskytem v textu a nikdy se nevyskytují bez kontextu. Můžeme je tak chápat jako pouhou vizuální pomůcku pro čtenáře, jelikož kocepce, kterou ukazují, je již postihnutá v textu. Za příklad nám může sloužit §48, kde je tématem řečová hra, jejíž smyslem je popsání barev na čtvercovém poli. Pravidla této hry jsou popsaná tak, že je možné si na jejich základě obrázky představit. Vzniká zde však prostor pro

omyly, které u obrázků hrozit nemusí (ne nutně). Horším příkladem je však §169, ve kterém Wittgenstein ukazuje, že psaný text určitým způsobem symbolizuje jazyk. Pokud čteme text, máme v hlavě znění čtených slov. To však neplatí v případě věty zapsané nesmyslným písmem (alespoň pro čtenáře), které Wittgenstein demonstruje následně:

> 169. But when we read, don't we feel the look of the words somehow causing our utterance? —— Read a sentence. — And now look along the following sequence
>
> 
>
> and utter a sentence as you do so. Can't one feel that in the first case

(ukázka symbolů v §169, obrázek převzatý ze[23])

Symboly, které Wittgenstein pro demonstraci používá, jsou už z principu strojově nezapsatelné. Bylo by sice možné je nahradit jinými znaky, které se vyskytují v unicode sadě, ale to by šlo přímo proti smyslu těchto symbolů. Nehledě na to je možné, že se tyto znaky (nebo spíše jejich tokeny), kterými bychom plánovali Wittgensteinovy symboly nahradit, nacházejí již v před-trénovaném modelu. Mohlo by to tak vést k zkreslení myšlenek obsažených v textu.[23]

Velikou roli může také hrát zaměnitelnost myšlenek. Jak už bylo popsáno v sekci B, text *Zkoumání* má dialogickou podobu. Wittgensteinův vypravěč popisuje teorii a Wittgensteinův tazatel vznáší námitky. Přičemž není jasné, jak na sebe jednotlivé promluvy reagují a kdy sledujeme čí perspektivu. Návodné nám v tomto ohledu mohou být dlouhé pomlčky „—“, které označují změnu perspektivy. I tak však není jasné, jaký Wittgensteinův hlas se zhošťuje slova. Jak bylo uvedeno ve stejné sekci této práce, panují také určité neshody kolem názoru, jak číst text *Zkoumání*. Které názory jsou opravdu Wittgensteinovy a které jsou zde jen pro to, aby umožnili průběh argumentace. Pokud tedy budeme brát text, jakožto souvislý řetězec znaků, znamená to, že se často vedle sebe objevují kontradiktorní názory, a to v rychlém sledu za sebou. Je tedy nutné míti na paměti, že některé naučené vzorce, nemusí být reprezentativní Wittgensteinových názorů.

Výsledný soubor textu *Filosofických* zkoumání, na kterém jsem prováděl trénování modelu, jsem zamýšlel uveřejnit na svém Github účtu. To především za účelem transparentnosti mého pracovního postupu. Dalším důvodem by bylo usnadnění práce (kterou si tento proces žádá) lidem, kteří by o tento očištěný text projevili zájem. To však bohužel není možné, jelikož se na text, i v této podobě, vztahuje autorský zákon. Není tak

pro mne možné jej distribuovat bez právních následků. Jedinou alternativou, která mě napadá je sestavení Python skriptu, který by tento proces tvorby trénovacího textu automatizoval. Zájemce o mojí verzi textu by dodal svojí originální pdf verzi *Filosofických zkoumání* a spustil poskytnutý Python skript. Pokud by skript fungoval správně, vznikla by verze *Zkoumání*, identická té mé.

2.3 Práce s jazykovým modelem GPT-NeoX

S modelem jsem pracoval za použití programovacího jazyka Python. To z prostého důvodu, že původní implementace tohoto jazykového modelu je právě v tomto jazyce. Konkrétně původní codebase byl vyvinutý a testovaný pro Python verze 3.8. Github repozitář modelu[9] uvádí, že se zdá, že model funguje také pro Python verze 3.9, ale jeho použití se nedoporučuje. Model pro své fungování dále využívá framework strojového učení PyTorch verze 1.8, nebo pozdější.

Při své práci jsem používal služby MetaCentra[10], což je česká virtuální organizace sdružení CESNET. Tato organizace je odpovědná za budování národního gridu a poskytuje tak distribuovanou výpočetní infrastrukturu, která se skládá z jednotlivých výpočetních center nacházejících se v České republice. Disponují tak dostatečným hardwarem a tedy i výpočetním výkonem pro účely mé práce. Tato výpočetní infrastruktura je navíc volně otevřená všem akademickým pracovníků a studentům vědeckovýzkumných institucí České republiky.

Při práci v MetaCentru je nutné využívat některý z dostupných singularity image kontejnerů. Jedná se o FOSS (free and open source) software, který provádí virtualizaci na úrovni operačního systému (kontejnerizaci). Singularity byl designovaný, aby splňoval nároky vysoce výkonných výpočetních systémů. U nich se totiž počítá s tím, že je bude využívat vysoké množství uživatelů, přičemž každý z nich by měl mít přístup pouze ke svým souborům. Singularity se tak spouští se stejnými privilegii uživatele a blokuje zvyšování pravomocí uvnitř kontejneru. Je možné si vytvořit vlastní kontejner image a nahrát ho do prostředí MetaCentra, což je užitečné pokud pro vaše účely potřebujete speciální sadu nástrojů. To však v mém případě nebylo třeba. MetaCentrum totiž poskytuje velikou sadu singularity imagů, z nichž několik je určených přímo pro deep learning. Po určité době strávené výběrem nejvhodnějšího kontejneru jsem se nakonec rozhodl pro

[9] https://github.com/EleutherAI/gpt-neox
[10] https://metavo.metacentrum.cz/

NGC (Nvidia GPU Cloud) PyTorch container, verze 21.05[11]. Ten obsahuje Python verze 3.8, který je nezbytný pro práci s před-trénovaným modelem, stejně jako PyTorch verze „1.9.0a0+2ecb2c7“. Během výběru vhodného Singularity kontejneru, jsem také musel věnovat dostatečnou pozornost, aby byla zachována kompatibilita mezi verzí softwaru CUDA a také verzí ovladačů GPU na stroji MetaCentra.

Při své práci jsem využíval Python knihovnu *transformers*[12] od společnosti Hugging Face. Jedná se o balíček, který uživateli poskytuje open-source implementaci vybraných transformerových modelů (jak pro práci s textem, tak obrazem i zvukem). Tato knihovna je zároveň kompatibilní s nejpoužívanějšími frameworky strojového učení PyTorch, TensorFlow a JAX. Za jejího využití je uživateli usnadněné stahování, trénování a používání volně dostupných před-trénovaných modelů.

Při své původní koncepci práce jsem zamýšlel využít původní kód poskytovaný společností EleutherAI na svém Githubu. Po několika pokusech o spuštění jsem však stále narážel na problémy s kompatibilitou. Stále si nejsem jistý, zda problém vznikal mou nedbalostí, či prostou nekompatibilitou kódu a poskytnutého prostředí MetaCentra. Poté co jsem však začal používat knihovnu *transformers*, přestaly se problémy vyskytovat. Přešel jsem tedy k této metodě.

V průběhu práce s modelem jsem založil repozitář (https://github.com/tvrzj/witt-gpt) na svém Github účtu, za účelem publikování kódu se kterým jsem pracoval. Repozitář by tak měl obsahovat Python skripty za pomoci kterých je možné: stáhnout před-trénovaný model GPT-NeoX, otestovat spustitelnost modelu, provádět trénování modelu a na závěr generovat text na dotrénovaném modelu. V repozitáři by se měl také nacházet bash skript za pomoci kterého jsem nastavoval prostředí v MetaCentru. Z důvodu velikosti natrénovaného modelu pro mě není možné jej taktéž začlenit do Github repozitáře. Z tohoto důvodu jsem se rozhodl jej sdílet na některém z dostupných cloudových úložišť a odkaz na stažení modelu vložit do repozitáře. Nezmiňuji konkrétní platformu, jelikož se situace ohledně dostupnosti cloudového úložiště může rapidně změnit. Po ukončení diplomové práce (této) jí taktéž hodlám přiložit do zmíněného repozitáře v podobě pdf souboru. Do Github repozitáře také plánuji začlenit ukázky textu, jaký je model schopný generovat.

---

[11] https://docs.nvidia.com/deeplearning/frameworks/pytorch-release-notes/rel_21-05.html#rel_21.05

[12] https://huggingface.co/docs/transformers/index

Bohužel v moment odevzdání této diplomové práce stále nemám natrénovaný model, který by dosahoval stupně natrénování, se kterým bych byl spokojený. Jelikož existence a odevzdání natrénovaného modelu není formální podmínkou pro úspěšné odevzdání diplomové práce, rozhodl jsem se s jeho publikací počkat do doby, kdy uznám za vhodné. Jednám tak dle svého nejlepšího uvážení a dle bodu devět *The Asilomar AI Principles*, publikovaných institutem Future of Life („Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications“). Nečiním si iluze, že by měl špatný stav mého jazykového modelu dalekosáhlé dopady na společnost, přesto bych nerad publikoval práci, která na to není připravená. Neočekávám, že by trénování jazykového modelu mělo zabrat více než několik dní práce navíc. V moment kdy bude práce na modelu hotová, bude uveřejněný na místech, které byly zmíněné v předchozím odstavci.[31]

Závěr

V první sekci této práce jsem nastínil původ myšlenek, které vedly vytvoření moderních systémů umělé inteligence. Sledoval jsem jejich evoluci od prvních pokusů o vytvoření inteligentního systému v podobě Perceptronu až po vznik transformerové architektury, která nyní platí za „state of the art“ v tomto oboru. V této sekci jsem také uvedl seznámení s tím, jak fungují základní koncepty oboru, důležité pro pochopení fungování těchto systémů. Dále jsem prozkoumal koncept modelování přirozeného jazyka za použití vektorové sémantiky. Tato kapitola tedy slouží, jakožto vhodný úvod do problematiky modelování přirozeného jazyka, pro osoby, které tyto metody doposud neznaly.

V druhé sekci této práce jsem provedl seznámení s Wittgensteinovým dílem. Uvedl jsem jakým způsobem se Wittgensteinovo dílo tradičně dělí. Provedl jsem krátkou argumentaci o způsobu čtení primárního textu jeho pozdní filosofie *Filosofická zkoumání*. Přidržoval jsem se při tom tak úvodu do studia tohoto textu od Davida Sterna *Wittgenstein's Philosophical investigations: an introduction*. Nastínil jsem také některé stěžejní koncepty, objevující se v text *Filosofická zkoumání*. Koncepty jako „rodinná podobnost“, „řečové hry“, „paradox soukromého jazyka“ a „paradox následování pravidla“. Tato část tedy jsouží, jakožto vhodný úvod do Wittgensteinova díla pro osoby, které se tímto dílem doposud nesetkali.

Ve třetí a finální sekci práce jsem popsal svůj postup, při trénování jazykového modelu na stěžejním textu Wittgensteinovi pozdní filosofie. Je zde také popsána má racionalizace výběru před-trénovaného jazykového modelu, který jsem použil. V rychlosti také popisuji práci v prostředí české gridové instituce, MetaCenta. Původně bylo mým záměrem poskytnou natrénovaný model souběžně s odevzdáním této práce. Bohužel, kvůli nespokojenosti se stupněm zpracování tak nečiním. Beru si za svůj osobní cíl model dotrénovat a následně jej publikovat na místech, která uvádím v práci. Vzniká zde také prostor, pro budoucí výzkum v tomto ohledu. Může se tak jednat o směr, kterým se vydám na svém budoucím studiu.

Resumé

The work is intended to serve as a bridge between two disciplines. The field of Natural Language Processing and the field of Philosophy. It is thus the culmination of my degree program in Philosophy for Artificial Intelligence. My goal is to describe the principles of natural language modeling to the extent that a layperson can begin to navigate the field. On the other hand, I also want to describe the ideas of Wittgenstein's late philosophy so that they can be understood by someone with no philosophical training. The third and final goal is to describe a procedure that could lead to the creation of a language model of a philosopher using pre-trained transformers.

In the first section of this paper, I outlined the origins of the ideas that led to the creation of modern artificial intelligence systems. I traced their evolution from the first attempts to create an intelligent system in the form of the Perceptron to the emergence of the transformer architecture, which is now considered the "state of the art" in this field. In this section, I have also provided an introduction to how the basic concepts of the field, important for understanding how these systems work, work. I also explored the concept of natural language modeling using vector semantics. Thus, this chapter serves as a convenient introduction to natural language modeling for people who have not been familiar with these methods before.

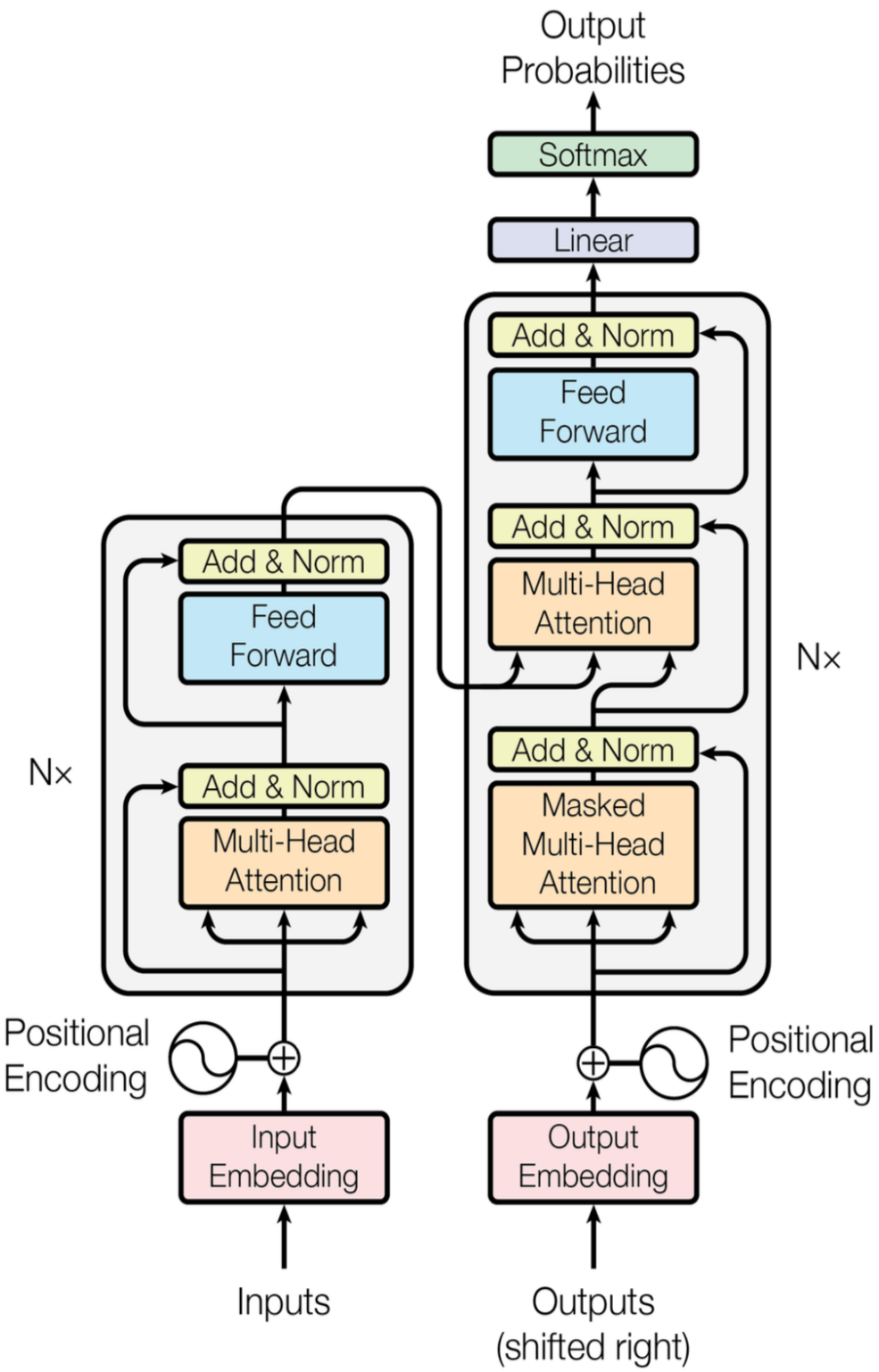
In the second section of this thesis, I have made an introduction to Wittgenstein's work. I have indicated how Wittgenstein's work has traditionally been divided. I have made a brief argument about the way in which the primary text of his late philosophy, Philosophical Investigations, is read. In doing so, I followed David Stern's introduction to the study of this text, Wittgenstein's Philosophical Investigations: an introduction. I have also outlined some of the key concepts appearing in the text Philosophical Investigations. Concepts such as 'family resemblance', 'speech games', 'the paradox of private language' and 'the paradox of rule-following'. This section thus serves as a suitable introduction to Wittgenstein's work for those who have not yet encountered it.

In the third and final section of the thesis, I describe my process of training the language model on the seminal text of Wittgenstein's late philosophy. My rationale for the choice of the pre-trained language model I used is also described. I also quickly describe my work in a Czech grid environment, MetaCent. It was originally my intention to provide the pre-trained model in parallel with the submission of this paper. Unfortunately, due to

dissatisfaction with the level of workmanship, I am not doing so. I take it as my personal goal to train the model and then publish it on the sites I mention in the paper. There is also room for future research in this regard. This may be the direction I take in my future studies.

Přílohy:

Diagram transformeru:



(obrázek převzatý ze[13])

Tabulka zastoupení dat v datasetu The Pile:

Část	Skutečná velikost	Váha	Epochy	Efektivní velikost	Průměrná velikost dokumentu
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
StackExchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19)	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB

Wikipedia (en)	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
Total				1254.20 GiB	5.91 KiB

Seznam literatury (dle pořadí výskytu):

[1] MITCHELL, Melanie. *Artificial intelligence: a guide for thinking humans*. New York: Farrar, Straus and Giroux, 2019. ISBN 978-0-374-25783-5.

[2] MCCARTHY, John. *Proposal for the Dartmouth Summer Research Project in Artificial Intelligence.* [online]. 1955. Dostupné z: https://web.archive.org/web/20070826230310/http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

[3] ROSENBLATT, Frank. *The Perceptron a Perceiving and Recognizing Automaton*. 1957. REPORT NO. 85-460-1

[4] ROSENBLATT, Frank. *The perceptron: A probabilistic model for information storage and organization in the brain*. *Psychological Review*. 1958, **65**(6), 386–408.

[5] MINSKY, Marvin a Seymour PAPERT. *Perceptron: an introduction to computational geometry.* 1969.

[6] BLACK, Sid, Stella BIDERMAN, Eric HALLAHAN, Quentin ANTHONY, Leo GAO, Laurence GOLDING, Horace HE, Connor LEAHY, Kyle MCDONELL, Jason PHANG, Michael PIELER, USVSN Sai PRASHANTH, Shivanshu PUROHIT, Laria REYNOLDS, Jonathan TOW, Ben WANG a Samuel WEINBACH. *GPT-NeoX-20B: An Open-Source Autoregressive Language Model* [online]. B.m.: arXiv. 14. duben 2022 [vid. 2023-05-10]. Dostupné z: http://arxiv.org/abs/2204.06745. arXiv:2204.06745 [cs]

[7] JURAFSKY, Daniel a James H. MARTIN. *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* [online]. Third Edition draft. 2023. Dostupné z: https://web.stanford.edu/~jurafsky/slp3/

[8] TAYLOR, Howard M. *An Introduction to Stochastic Modeling*. Fourth Edition. 2010. ISBN 978-0-12-381416-6.

[9] PAL, Saurabh. *Implementing Word2Vec in Tensorflow* [online]. [vid. 2023-08-21]. Dostupné z: https://medium.com/analytics-vidhya/implementing-word2vec-in-tensorflow-44f93cf2665f

[10] BROWN, Tom B., Benjamin MANN, Nick RYDER, Melanie SUBBIAH, Jared KAPLAN, Prafulla DHARIWAL, Arvind NEELAKANTAN, Pranav SHYAM, Girish SASTRY, Amanda ASKELL, Sandhini AGARWAL, Ariel HERBERT-VOSS, Gretchen KRUEGER, Tom HENIGHAN, Rewon CHILD, Aditya RAMESH, Daniel M. ZIEGLER, Jeffrey WU, Clemens WINTER, Christopher HESSE, Mark CHEN, Eric SIGLER, Mateusz LITWIN, Scott GRAY, Benjamin CHESS, Jack CLARK, Christopher BERNER, Sam MCCANDLISH, Alec RADFORD, Ilya SUTSKEVER a Dario AMODEI. *Language Models are Few-Shot Learners* [online]. B.m.: arXiv. 22. červenec 2020 [vid. 2023-05-29]. Dostupné z: http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs]

[11] CHUNG, Hyung Won, Le HOU, Shayne LONGPRE, Barret ZOPH, Yi TAY, William FEDUS, Yunxuan LI, Xuezhi WANG, Mostafa DEHGHANI, Siddhartha BRAHMA, Albert WEBSON, Shixiang Shane GU, Zhuyun DAI, Mirac SUZGUN, Xinyun CHEN, Aakanksha CHOWDHERY, Alex CASTRO-ROS, Marie PELLAT, Kevin ROBINSON, Dasha VALTER, Sharan NARANG, Gaurav MISHRA, Adams YU, Vincent ZHAO, Yanping HUANG, Andrew DAI, Hongkun YU, Slav PETROV, Ed H. CHI, Jeff DEAN, Jacob DEVLIN, Adam ROBERTS, Denny ZHOU, Quoc V. LE a Jason WEI. *Scaling Instruction-Finetuned Language Models* [online]. B.m.: arXiv. 6. prosinec 2022 [vid. 2023-08-07]. Dostupné z: http://arxiv.org/abs/2210.11416. arXiv:2210.11416 [cs]

[12] TOUVRON, Hugo, Thibaut LAVRIL, Gautier IZACARD, Xavier MARTINET, Marie-Anne LACHAUX, Timothée LACROIX, Baptiste ROZIÈRE, Naman GOYAL, Eric HAMBRO, Faisal AZHAR, Aurelien RODRIGUEZ, Armand JOULIN, Edouard GRAVE a Guillaume LAMPLE. *LLaMA: Open and Efficient Foundation Language Models* [online]. B.m.: arXiv. 27. únor 2023 [vid. 2023-05-10]. Dostupné z: http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs]

[13] VASWANI, Ashish, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N. GOMEZ, Lukasz KAISER a Illia POLOSUKHIN. *Attention Is All You Need* [online]. B.m.: arXiv. 5. prosinec 2017 [vid. 2023-05-10]. Dostupné z: http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs]

[14] *Neural machine translation with a Transformer and Keras* [online]. 20.8. Dostupné z: https://www.tensorflow.org/text/tutorials/transformer

[15] BENGIO, Yoshua, Réjean DUCHARME, Pascal VINCENT a Christian JANVIN. *A Neural Probabilistic Language Model*. *J. Mach. Learn. Res.* 2003, **3**(null), 1137–1155. ISSN 1532-4435.

[16] MANNING, Christopher D. *Human Language Understanding & Reasoning*. *Daedalus* [online]. 2022, **151**(2), 127–138 [vid. 2023-08-07]. ISSN 0011-5266, 1548-6192. Dostupné z: doi:10.1162/daed_a_01905

[17] DEVLIN, Jacob, Ming-Wei CHANG, Kenton LEE a Kristina TOUTANOVA. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* [online]. B.m.: arXiv. 24. květen 2019 [vid. 2023-05-29]. Dostupné z: http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs]

[18] SCHWITZGEBEL, Eric, David SCHWITZGEBEL a Anna STRASSER. *Creating a Large Language Model of a Philosopher* [online]. B.m.: arXiv. 9. květen 2023 [vid. 2023-06-29]. Dostupné z: http://arxiv.org/abs/2302.01339. arXiv:2302.01339 [cs]

[19] WITTGENSTEIN, Ludwig. *Tractatus Logico-Philosophicus* [online]. SIDE-BY-SIDE-BY-SIDE EDITION. nedatováno. Dostupné z: http://people.umass.edu/klement/tlp/

[20] WITTGENSTEIN, Ludwig, Marjorie PERLOFF a Ludwig WITTGENSTEIN. *Private notebooks: 1914-1916*. First edition. New York, NY: Liveright Publishing Corporation, a Division of W. W. Norton & Company, Independent Publishers Since 1923, 2022. ISBN 978-1-324-09080-9.

[21] STERN, David G. *Wittgenstein's Philosophical investigations: an introduction*. New York: Cambridge University Press, 2004. Cambridge introductions to key philosophical texts. ISBN 978-0-521-81442-3.

[22] BILETZKI, Anat a Anat MATAR. *Ludwig Wittgenstein*. *Stanford Encyclopedia of Philosophy* [online]. [vid. 2023-08-20]. Dostupné z: https://plato.stanford.edu/entries/wittgenstein/

[23] WITTGENSTEIN, Ludwig, G. E. M. ANSCOMBE, P. M. S. HACKER a Joachim SCHULTE. *Philosophische Untersuchungen =: Philosophical investigations*. Rev. 4th ed. Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell, 2009. ISBN 978-1-4051-5928-9.

[24] WITTGENSTEIN, Ludwig, Gertrude Elizabeth Margaret ANSCOMBE a Georg Henrik von WRIGHT. *Zettel*. Bilingual ed. Berkeley Los Angeles: University of California Press, 2007. ISBN 978-0-520-25244-8.

[25] GLOCK, Hans-Johann. *A Wittgenstein dictionary*. Oxford, OX, UK ; Cambridge, Mass., USA: Blackwell Reference, 1996. The Blackwell philosopher dictionaries. ISBN 978-0-631-18112-5.

[26] BUCUR, Anca a Sergiu NISIOI. *A Visual Representation of Wittgenstein's Tractatus Logico-Philosophicus* [online]. B.m.: arXiv. 13. březen 2017 [vid. 2023-08-20]. Dostupné z: http://arxiv.org/abs/1703.04336. arXiv:1703.04336 [cs]

[27] MALEČKOVÁ, Dita a Jan TYL. *Digitální filosof* [online]. [vid. 2023-08-20]. Dostupné z: https://digitalnifilosof.cz/

[28] BENDER, Emily M., Timnit GEBRU, Angelina MCMILLAN-MAJOR a Shmargaret SHMITCHELL. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜. In: *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* [online]. Virtual Event Canada: ACM, 2021, s. 610–623 [vid. 2023-08-19]. ISBN 978-1-4503-8309-7. Dostupné z: doi:10.1145/3442188.3445922

[29] BIDERMAN, Stella, Kieran BICHENO a Leo GAO. *Datasheet for the Pile* [online]. B.m.: arXiv. 13. leden 2022 [vid. 2023-05-29]. Dostupné z: http://arxiv.org/abs/2201.07311. arXiv:2201.07311 [cs]

[30] GAO, Leo, Stella BIDERMAN, Sid BLACK, Laurence GOLDING, Travis HOPPE, Charles FOSTER, Jason PHANG, Horace HE, Anish THITE, Noa NABESHIMA, Shawn PRESSER a Connor LEAHY. *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* [online]. B.m.: arXiv. 31. prosinec 2020 [vid. 2023-08-19]. Dostupné z: http://arxiv.org/abs/2101.00027. arXiv:2101.00027 [cs]

[31] Future of Life Institute. AI Principles. *Future of Life Institute* [online]. [vid. 2023-08-20]. Dostupné z: https://futureoflife.org/open-letter/ai-principles/