SOUČASNÉ TRENDY VE ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA

Datum konání: 5. 4. 2019

Místo konání: Arna Nováka 1, Brno, učebna D51

Název přednášky: Současné trendy ve zpracování přirozeného jazyka

Přednášející: Ing. Michal Hradiš, Ph.D.

Počet účastníků: 14

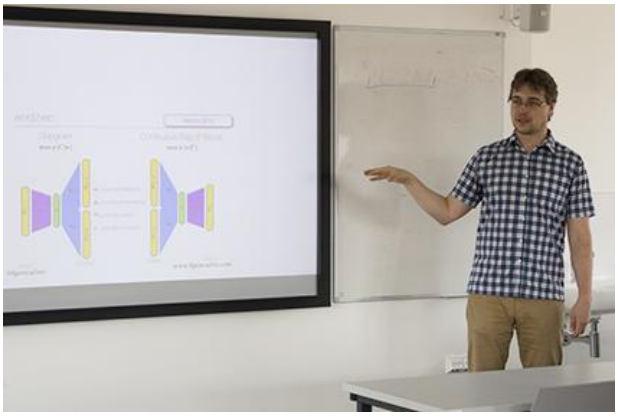
Zpracovala: Čechlovská Hana

Ing. Michal Hradiš, Ph.D., pracuje v Ústavu počítačové grafiky a multimédií Fakulty informačních technologií VUT Brno. Jeho náplní práce jsou systémy pro přepis historických tištěných i ručně psaných dokumentů a další zpracování jejich obsahu. Dlouhodobě se věnuje strojovému učení a počítačovému vidění. V minulosti se podílel na vyvíjení rychlé detekce objektů v obraze, zpracování dohledových záznamů po přirozená uživatelská rozhraní, zlepšování kvality obrazu a kontroly kvality výrobků. V současné době se zaměřuje převážně na využití hlubokých neuronových sítí pro obraz a text. Říká, že nemá rád slova, pracuje pouze s písmeny.

Na první dubnovou přednášku z jarního cyklu odborných přednášek a exkurzí jsme všichni přišli s očekáváním, že se dozvíme zajímavosti a novinky ze světa zpracování přirozeného jazyka. Informační leták sliboval, že budou představeny reprezentace slov, jazykové modely, analýza sentimentu, strojový překlad a základní myšlenky stojící za některými současnými trendy využití strojového učení ve zpracování přirozeného jazyka. S radostí můžu říct, že pan doktor stihl zmínit víc, než bylo slíbeno.

Úvodem proběhlo krátké představení pana doktora Hradiše a jeho práce. Pro zjištění míry našich znalostí se nejdříve zeptal, co můžeme zkoumat na jazyku. Ozývalo se třeba parsing, tagging, jazykové modely, koreference, pojmenování entit, inference a počítačové

porozumění textu. Po objasnění a rozvedení pojmů do větší hloubky, zmínil i analýzu GLUE benchmark, což je soubor zdrojů pro trénování, ohodnocování a analýzu přirozeného jazyka systémy pro porozumění. Další zajímavostí bylo vyvinutí velkého datasetu otázek pro server Quora. Aby uživatel věděl, jestli jeho dotaz byl již zodpovězen, systém porovná všechny otázky s nově zadanou a uživatele buď rovnou přesměruje na odpověď, nebo ho nechá položit otázku. Úplnou novinkou pro mě bylo Paragraph-Level Question-Answer Pairs, kdy se algoritmus snaží zjistit odpověď na zadanou otázku z určitého odstavce.



V další části nám představil aplikaci ROI Hunter Easy, na jejímž vývoji se podílí. Cílem systému je generování krátkého textu z oficiálního popisu produktu, který zaujme zákazníka natolik, že produkt koupí. Důležitou otázkou je, na základě čeho by si člověk mohl daný výrobek koupit. Je důležitější, jestli je kabelka modrá, nebo že je kožená? Potřebuje k tomu obrázek, nebo stačí text?

Druhou polovinu přednášky pan doktor Hradiš věnoval oblasti NLP. Na otázku, jak reprezentujeme slova, měl jednoznačnou odpověď – word embedding, aneb číselná reprezentace slov. Princip spočívá v převedení slov na v mnoharozměrném prostoru tak, aby se slova s podobnými vlastnostmi nacházela blízko sebe. Jako příklad můžeme uvést

Francie – Paříž + Bratislava = Slovensko.

Přínosné bylo i představení knihovny fastText, která pracuje s n-gramy písmen reprezentovanými vektory. Ty se sečtou, vygenerují slovo a následně celkový vektor slova. Knihovna má zarovnání pro 44 jazyků.

Závěrem nás téměř vyškolil v jazykových modelech, pravděpodobnostech slov, zběžně zmínil překládání z obrázku do textu, např. Image2Seq, kdy je výstupem popis obrázku, a nástroj BERT, jehož úkolem je predikce další věty. Bidirectional Encoder Representations from Transformers je výpočetně hodně složitý, ale momentálně nejlepší nástroj pro práci na datech, který oproti ostatním nástrojům dokáže poskytnout i těžší výsledky doplňování.

Pan doktor Michal Hradiš byl velmi dobrý přednášející, který dokázal udržet naši pozornost do samého konce. Všichni zúčastnění jsme odcházeli s hlavou plnou hodnotných informací, které jsou užitečné nejen pro studenty počítačové lingvistiky, ale i pro „normální“ lingvisty. Přednášku bych rozhodně doporučila a doufám, že na Masarykově univerzitě nepřednášel naposledy.