Specifikace ročníkového projektu

Vojtěch Švandelík

12. října 2020

V následujícím dokumentu je uvedena specifikace ročníkového projektu na *Matematicko-fyzikální fakultě Univerzity Karlovy*. Vedoucím ročníkového projektu je Mgr. Martin Popel, Ph.D. z Ústavu formální a aplikované lingvistiky.

1 Cíl projektu

Obsahem ročníkového projektu je identifikovat chyby v automatickém překladu neuronového překladače *CUBBITT* a navrhnout způsoby jejich řešení.

2 Popis fází

2.1 Identifikace chyb

Při hledání chyb se bude vycházet z korpusu *CzEng 2.0*[1]. Tento obsahuje české i anglické texty přeložené prostřednictvím výše zmíněného překladače do opačného jazyka. V přeložených textech proběhne poloautomatické filtrování, které se pokusí potvrdit již dříve identifikované překladové problémy a rovněž se zaměří na hledání nových.

Prozatím známé (a vcelku významné) překladové chyby jsou nevhodné překládání vlastních jmen osob a překlad jednotek u čísel, kdy hodnota ztrácí faktický význam.

2.2 Hledání řešení

Prvotní fází napravování překladových chyb bude vkládání do textu speciální unicode znaky. Následně se bude zkoumat, zda-li překladač obalené části vět bude či nebude překládat. Dle získaných výsledků budou následovat další kroky. Jedním z nich by mohlo být přetrénování testovacích dat s ohledem na zamezení překládání speciálně označených tokenů.

2.3 Vyhodnocování výsledků

Úspěšnost vylepšení překladu bude vyhodnocována pomocí dostupných nástrojů – např. metrika Bleu.

3 Výstup projektu

Výstupem ročníkového projektu by měla být rozvaha nad nalezenými překladovými chybami, společně s rešerší týkající se jejich oprav.

V rámci implementační části by měl vzniknout základ evaluace výsledků a především by mělo dojít k využití alespoň některých návrhů řešení překladových chyb.

Výstupy by měly posloužit jako podklady pro psaní bakalářské práce.

[1] http://ufal.mff.cuni.cz/czeng