Specifikace ročníkového projektu

Vojtěch Švandelík

29. března 2021

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 *LINDAT Translation*[1] a navrhnout způsoby jejich řešení. Konkrétně se bude jednat o překlad mezi českým a anglickým jazykem v obou směrech.

2 Výstup projektu

Výstupem ročníkového projektu bude balíček v jazyce Python, který bude možné volat na jednotlivé již přeložené věty a který bude vracet jejich „opravený“ překlad.

K tomuto balíčku vzniknou dva nástroje pro jeho otestování:

- `CLI` program v jazyce Python, který umožní spustit opravu vět ze standardního vstupu a opravené věty bude vypisovat na standardní výstup (tento bude použit pro opravu trénovacích dat překladače LINDAT);
- `WSGI` webová aplikace využívající standardní webové technologie na straně frontendu a Python na straně backendu pro uživatelsky přívětivé otestování.

Výstupy ročníkového projektu poslouží jako podklady pro tvorbu bakalářské práce.

3 Opravované jevy

Primárním cílem oprav budou číselné údaje s běžně užívanými jednotkami. Při překladu totiž u některých vět dochází k některým z následujících jevů:

a. překladač ponechá původní číselný údaj ale použije jinou jednotku;

b. překladač „přeloží“ i číselný údaj i jednotku;

c. překladač nechá beze změny původní číselný údaj obsahující oddělovač tisíců či desetinných čísel, čímž se ztratí vypovídající hodnota.

Oprava chyb se bude týkat nejběžněji používaných fyzikálních jednotek, tj. jednotky délky, hmotnosti, teploty, rychlosti a rovněž mezinárodních měn (CZK, USD, GBP, EUR).

[1] https://lindat.mff.cuni.cz/services/translation/

Tabulka 1: Ukázka základních překladových chyb, jejichž opravu bude vytvořený nástroj podporovat.

Zdrojová věta	Přeložená věta	Oprava
Veronika Stýblová vážila o 20 kilo víc.	Veronica Bean weighed 20 pounds more.	Veronica Bean weighed 20 kilograms more.
Je vysoký pouhých 190 cm.	He's only six feet tall.	He's only 190 cm tall. *
Suchou váhu udává výrobce 179,5 kg.	The dry weight is given by the manufacturer at 179,5 kg.	The dry weight is given by the manufacturer at 179.5 kg. **

** Překladač se pokusil číselnou hodnotu přepočítat, ale výsledný číselný údaj v kontextu věty ztratil část informace. Původních 190 centimetrů se totiž přeložilo jako šest stop, což odpovídá 183 centimetrům.*
*** Číselná hodnota 179 a půl kilogramu byla překladačem ponecháním desetinné čárky přeložena jako 179 tisíc kilogramů.*

4 Použité metody opravy chyb

V rámci ročníkového projektu budou chyby nalézány a opravovány prostřednictvím nástroje, který vznikne v programovacím jazyce Python (verze 3.8), který tak bude činit na základě jazykových heuristik. Tyto budou zpřesňovány pomocí externích nástrojů – především nástroje pro word-alignment. Ten bude nápomocen při opravování vět, kde se nachází více číselných údajů s jednotkami a tedy umožní identifikovat k sobě patřící dvojice. Jako součást nástroje vznikne rovněž vestavěný word-aligner, který bude využívat naivní implementaci, kdy k sobě přiřadí dvojice čísel a jednotek dle pořadí ve větě.

Pro vývoj nástroje, analyzování překladových chyb a měření úspěšnosti bude využit česko-anglický paralelní korpus *CzEng 2.0*,[2] který byl sestaven na Ústavu formální a aplikované lingvistiky MFF UK. Tento kromě autentických trénovacích dat, využívaných pro trénovaní modelu překladače, obsahuje již syntetická data přeložená překladačem. Tyto jsou vhodná pro odhalování překladových chyb.

Tabulka 2: Statistika počtu pro nástroj zajímavých vět ve vzorku 100 000 vět z syntetických česko-anglických dat korpusu *CzEng 2.0*.

	Počet		
Typ věty	**Fyzikální jednotky**	**Měny**	**Celkově**
věta s alespoň jedním číslem s jednotkou	603	591	1194
vět, kde byly zachovány všechna čísla s jednotkami	565	438	1003
věta, kde překladač nezachoval alespoň jedno číslo	0	3	3
věta, kde překladač nezachoval alespoň jednu jednotku	16	133	149
věta, kde byl nevhodně přeložen oddělovač tisíců / desetinných čísel u čísla s jednotkou	7	4	11

Pro metody opravy chyb je rovněž podstatné, zdali je ve větě uveden pouze jeden číselný údaj s jednotkou, anebo zdali jich je více. V případě první varianty je totiž oprava chyb typicky zjednodušena (není třeba word-alignment).

Tabulka 3: Statistika počtu čísel s jednotkami ve větách ve vzorku 100 000 vět z syntetických dat česko-anglických korpusu *CzEng 2.0*.

Typ věty	Počet
věta s jedním číslem s jednotkou	1009
věta s více čísly s jednotkami	185

[2]Tom Kocmi, Martin Popel a Ondřej Bojar. „Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords". In: *arXiv preprint arXiv:2007.03006* (2020).

5 Parametrizovatelnost nástroje

Funkci nástroje bude možné ovlivnit prostřednictvím několika parametrů. Tyto budou založeny hlavně na různých režimech opravy chyb v překladech.

Nástroj bude mít dva hlavní režimy. V prvním režimu, který bude využit pro trénování modelu překladače *LINDAT Translation*, bude nástroj hledat překladové chyby a tyto se pokusí opravit (provede se co nejmenší možná změna tak, aby se chyba odstranila).

V druhém režimu půjde o především webový nástroj, který umožní uživatelům upravit výstup překladače tak, aby využíval jednotky zvolené uživatelem.

Oba režimy fungování budou ovlivnitelné dvěma číselnými konstantami:

- míra tolerance odchylky překladu u číselných údajů u všech překladových chyb;
- míra tolerance odchylky překladu u číselných údajů, které budou považovány za přibližné (tj. před číselnou hodnotou se bude vyskytovat typu „asi“, „přibližně“, atp.).

Dalšími parametry, které bude možné změnit v konfiguraci nástroje, bude použití online word-aligneru a rovněž režim získání kurzu pro přepočet měn (bude možné použít fixní kurz či získat aktuální kurz z webového portálu *České národní banky*).

6 Technické parametry

Nástroj bude vyvíjen v jazyce Python ve verzi 3.8. Budou využity standardní knihovny. K nástroji vznikne technická i uživatelská dokumentace. Technická dokumentace bude především generována ze zdrojových kódů prostřednictvím nástroje *Sphinx* a bude doplněna popisem architektury a zvolených technologií. K nejdůležitějším částem aplikace vzniknou unit-testy ověřující správnou funkci.

S externím nástrojem – word-alignerem bude aplikace komunikovat prostřednictvím standardního komunikačního protokolu *HTTP*.

Aplikace bude využívat verzovací systém *Git*.

7 Ověřování funkčnosti aplikace

Kromě výše zmíněných unit-testů vznikne tzv. „dev-set“, tj. množina ručně vybraných vět s jejich správnými překlady. Na této množině vět bude aplikace spouštěna a bude měřena její úspěšnost prostřednictvím běžně užívaných statistických nástrojů (především úspěšnost počtu správně opravených vět).

8 Licence

Aplikace bude publikována prostřednictvím služby *GitHub*[3] pod licencí MIT.

[3] `https://github.com/vsvandelik/lindat-translation-postprocessor`