Paralelní korpus z Wikipedie

Adéla Štromajerová

Zadání práce

Cílem práce je sestavit paralelní korpus na základě takových editací Wikipedie, které prokazatelně přebírají a jsou překlady textů z jiných jazykových mutací Wikipedie. Často se překládají i celé stránky z anglické do české Wikipedie a v takových případech musí být toto převzetí správně indikováno. Vyžaduje to jak licence Wikipedie, tak editorská kultura.

Úkoly. Student se seznámí se strukturou databáze Wikipedie, zejména se systémem pro správu verzí stránek. Pomocí skriptů extrahuje všechny takové editace, které jsou prokazatelně překlady a vytvoří z nich zarovnané texty. Hlavní jazykový pár bude angličtina a čeština, nicméně skripty by měly umět extrahovat paralelní data i pro jiné jazykové páry. Takto extrahované texty následně student upraví do podoby zarovnaných vertikálních souborů, ze kterých se vytvoří paralelní korpus pro nástroj Sketch Engine, který se vyvíjí ve spolupráci s Centrem pro zpracování přirozeného jazyka na Fakultě informatiky. Tímto se výsledný korpus zpřístupní uživatelům na Masarykově univerzitě.

Práce bude psána anglicky.

Současný stav problematiky

Korpusová lingvistika je v dnešní době zajímavým a slibným oborem. Korpusy jsou obecně považovány za velmi cenné nástroje a jsou hojně používány např. na poli zpracování přirozeného jazyka (NLP – natural language processing) jako nedocenitelné zdroje dat. V určitých podoborech NLP se používají specifické druhy korpusů. Pro statistický strojový překlad (SMT – statistical machine translation) jsou například nezbytné korpusy paralelní.

Paralelní korpus se skládá ze dvou nebo více korpusů v různých jazycích. Tyto korpusy obsahují buď texty, které byly přeloženy z jednoho jazyka do druhého, nebo texty, které byly vytvořeny simultánně ve více jazycích [1]. Používanější jsou korpusy obsahující simultánně vytvořené texty. Tyto texty obvykle pocházejí z komunikace ve vícejazyčných komunitách, např. z Organizace spojených národů, Evropské unie nebo Kanady [2]. Otevřený paralelní korpus OPUS například obsahuje paralelní korpus Europarl, který se skládá z textů, jež byly pořízeny ve 21 jazycích při jednáních v Evropském parlamentu [3]. Další paralelní korpus, The Canadian Hansard Corpus, sestává z debat kanadského parlamentu, které byly uveřejněny v úředních jazycích země, v angličtině a francouzštině [4].

Takovéto korpusy lze použít v mnoha oborech. Používají je překladatelé nebo např. učitelé a jejich studenti k vyhledání určitých výrazů, zjištění, jak se tyto výrazy používají, a ke zkoumání rozdílů mezi jazyky [1]. Hojně se také používají pro SMT, kde slouží jako zdroje trénovacích dat. Jelikož je SMT stále důležitější, nabývají na důležitosti i paralelní korpusy, bez kterých by SMT nemohl fungovat a nemohl by se dále vyvíjet. Výše zmíněný Europarl byl například vytvořen přímo pro potřeby SMT [3]. Paralelní

korpusy se dále používají na poli lexikografie, kde tím, že poskytují reálné, neumělé příklady jazyka, pomáhají k tomu, aby slovníky lépe odrážely stav jazyka a jeho použití [5].

K vytvoření paralelního korpusu je potřeba velké množství dat. Takováto data se v dnešní době nacházejí na Internetu. Jedno z míst, kde je možné najít texty na stejné téma ve více jazycích, je Wikipedie, otevřená webová encyklopedie, která nabízí články ze všech možných oborů ve všech možných jazycích. Wikipedie rychle roste, jelikož přispívat může v podstatě každý a mechanismus úprav a tvorby článků je jednoduchý a intuitivní [6]. Wikipedie proto obsahuje texty, které nemusí sloužit pouze jako informativní články, ale také jako data pro mnoho aplikací.

V minulosti již bylo z Wikipedie vytvořeno několik korpusů. Zmínit můžeme např. The Wikipedia Corpus, který obsahuje texty z anglické Wikipedie [7]. Existuje také několik paralelní korpusů pocházejících z Wikipedie, např. čínsko-japonský paralelní korpus vytvořený za účelem zlepšení SMT mezi těmito dvěma jazyky [8].

Před samotným vytvářením paralelního korpusu z Wikipedie je třeba identifikovat, které dvojice článků jsou překlady. Toto je jednoduché, jelikož všechny přeložené články na Wikipedii musí obsahovat formuli, která je jako přeložené označuje, a přeložený článek také musí obsahovat odkaz na článek původní [9]. Překlady na Wikipedii jsou většinou z angličtiny do ostatních jazyků, ačkoli samozřejmě existují články, specifické pro určitou zemi, které jsou poté, co témata v nich obsažené nabudou na důležitosti, přeloženy do angličtiny, popř. jiných jazyků.

Česko-anglický pár je obsažen v paralelním korpusu Europarl. Nicméně, pokrytí tohoto jazykového páru by mohlo být lepší. Vytvořením paralelního korpusu z českých a anglických verzí téhož článku na Wikipedii můžeme pomoci zlepšit SMT mezi těmito dvěma jazyky a také poskytnout další zdroje znalostí lexikografům, překladatelům a učitelům.

Odkazy

[1] Susan Hunston. *Introduction to a corpus in use*. Cambridge Press, 2002.

[2] *Parallel corpora*. URL: http://www.ilc.cnr.it/EAGLES/corpustyp/node20.html.

[3] *Europarl Parallel Corpus*. URL: http://www.statmt.org/europarl/.

[4] *The Canadian Hansard Corpus*. URL: http://spraakbanken.gu.se/lb/pedant/parabank/node6.html.

[5] David Lindemann. "Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair". In: *Procedia – Social and Behavioral Sciences* 95 (2013), s. 249–257. DOI: http://dx.doi.org/10.1016/j.sbspro.2013.10.645.

[6] Wikipedia. *Wikipedia*. URL: https://en.wikipedia.org/wiki/Wikipedia.

[7] *The Wikipedia Corpus*. URL: http://corpus.byu.edu/wiki/.

[8] Chenchui Chui, Toshiaki Nakazawa a Sadao Kurohashi. *Constructing a Chinese-Japanese Parallel Corpus from Wikipedia*. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/21_Paper.pdf.

[9] Wikipedia. *Translation*. URL: https://en.wikipedia.org/wiki/Wikipedia:Translation.

Rejstřík