Jindřich Libovický, Rudolf Rosa, Josef Čech a další

Softwarový projekt *FilmTit*
Specifikace

Počet řešitelů: 5 až 6

Úvod

Překlad filmových a seriálových titulků je činnost, které se ve svém volném čase věnuje poměrně velké množství lidí. Přestože v komerčním světě je již mnoho let dostupné relativně velké množství aplikací, které usnadňují lidmi prováděný překlad, pro překlad titulků dobrovolníky takový nástroj dosud chybí. Přitom právě nekomerční a komunitní charakter překládání titulků a obrovské (a stále rostoucí) množství dostupných dat se přímo nabízí pro vývoj takového nástroje.

Titulky se zároveň jeví jako velmi dobrý zdroj paralelních dat. Bývají používány při tvorbě paralelních korpusů, využitelných pro statistický strojový překlad (např. *CzEng* http://ufal.mff.cuni.cz/czeng/). Ve filmovém průmyslu existují i pokusy o strojový překlad titulků, které by potom překladatel jen částečně poupravil.[1]

Překladové paměti jsou v současnosti nejrozšířenějším nástrojem pro usnadnění překladu. Většinou se z důvodu výkonu osobních počítačů a snaze „neznečistit“ záznamy různým překladem stejných termínů z různých oblastí udrží co nejmenší, zaměřené jen na úzkou doménu. V případě filmových titulků takové nejednoznačnosti nehrozí. Naopak věříme, že podobnost scénářů filmů povede k tomu, že s rostoucím objemem dat bude překladová paměť poskytovat stále lepší výsledky.

Cílem našeho projektu je vytvořit aplikaci, která bude usnadňovat překlad filmových a seriálových titulků. Jádrem aplikace bude rozsáhlá překladová paměť, která se bude používáním dále rozšiřovat a obohacovat, a která bude existovat pouze v jedné veřejně přístupné instanci.

V naší práci se budeme soustředit na jazykový pár angličtina – čeština, jádro aplikace by ale mělo být na zvolených jazycích nezávislé.

Aplikace

Aplikaci bude mít dvě části: serverová část bude tvořena překladovou pamětí, klientská část pak bude uživatelům nabízet překladatelské rozhraní. Obě části spolu budou komunikovat vhodným způsobem (nejspíše JSON).

---

[1] Flanagan, Marian. *Using Example-Based Machine Translation to translate DVD Subtitles.* Proceedings of the 3rd Workshop on Example Based Machine Translation. Dublin, Ireland. 2009.

Překladatelské rozhraní bude především poskytovat přirozené a praktické prostředí pro práci samotných překladatelů a zprostředkovávat komunikaci s překladovou pamětí. Základem je samozřejmě načtení anglických titulků a podpora jejich překladu do češtiny, s využitím návrhů z překladové paměti. Pro zvýšení uživatelského komfortu bude aplikace umožňovat přehrávání příslušného filmu (bude-li jej překladatel mít uložený na svém disku) synchronizované s překládanými titulky. Přidruženy budou i další funkce usnadňující překlad, například nástroj pro úpravu časování titulek nebo, bude-li to možné, integrace strojového překladu některým již existujícím překladačem.[1] Rozhraní bude vytvořeno v podobě webové aplikace.

Překladová paměť bude přijímat požadavky klientské aplikace na překlad jednotlivých úseků titulků, vyhledá v databázi podobné známé texty a několik *nejlepších* z nich odešle zpět. Pro ohodnocení se pokusíme kromě běžných metrik shody využívat i zpětnou vazbu od uživatelů, kteří budou mít možnost hodnotit výstupy překladové paměti. Ověříme také možnost využití některých meta informací, jako je shoda filmové či seriálové řady, shoda filmového žánru apod. Zároveň bude překladová paměť samozřejmě přijímat veškeré překlady vytvořené uživateli a ukládat si je do databáze. Celkově by tedy mělo docházet k neustálému zkvalitňování obsahu překladové paměti. Překladová paměť bude napsána v jazyce Java a bude vybudována nad vhodným existujícím databázovým systémem.

Data

Práce s filmovými titulky se od práce s obecnými texty v mnohém liší. Věříme, že většina odlišností bude představovat spíše výhody (například omezená doména nebo velké množství dat využitelných pro počáteční naplnění překladové paměti), jistě tomu tak ale nebude ve všech případech.

V první řadě je nutné stanovit vhodnou základní jednotku textu, která bude celistvě ukládána do překladové paměti. Je obvyklé používat párování na úrovni vět, nicméně jednotlivé úseky titulků často dělí věty na více částí a/nebo sdružují více vět dohromady, k čemuž lze přistupovat jako k šumu, který je třeba odstranit, ale i jako k informaci navíc, kterou lze využít.

Pravděpodobně bude možné zlepšit výsledky preprocessingem pojmenovaných entit (jména, čísla apod.), případně i jejich odděleným překladem (přičemž data obsažená v překladové paměti je v tomto případě možné využít pro vybudování jazykového modelu).

Filmové titulky také umožňují využití mnoha metadat, která lze získat pomocí API IMDb.com (International Movie Database) nebo z její otevřené obdoby TMDb.org (The Movie Database). Pokusíme se využít alespoň žánr filmu (a v návrzích překladu upřednostnit ty z žánrově podobných filmů), případně i další informace, jako jsou jména postav ve filmu (zde se samozřejmě nabízí propojení s preprocessingem pojmenovaných entit).

Jako zdroj dat pro počáteční naplnění překladové paměti budou využity filmové a seriálové titulky volně dostupné na internetu. Například na serveru opensubtitles.org, který je často používán pro podobné účely, je v současné chvíli k dispozici více než 65 000 filmových titulků

[1] Původně zamýšlený Google Translate se bohužel stává placeným, možnosti zapojení strojového překladu proto musíme přehodnotit.

v češtině. Potřebné spárování titulků je navíc snazší než párování obecných textů (zejména díky specifické struktuře titulků) a existují pro něj efektivní metody.[1]

Závěr

Výsledkem by měl být komplexní systém usnadňující dobrovolným překladatelům překlad filmových a seriálových titulků. Uživatelé budou navíc svou prací systém dále vylepšovat.

Kromě svého původního využití bude možné systém s dobře strukturovaně uloženými titulky snadno použít k vytvoření paralelního korpusu a k výzkumu jazyka používaného v titulcích.

[1] Itamar, Einav – Ita, Alon. *Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora*. Proceedings of the 6th international conference on Language Resources and Evaluation. 2008.