

Zadanie softwarového projektu Translation web editor

Zadávatel'

David Čaněk
MemSource, s.r.o., Praha
<http://www.memsource.com>

Vedúci projektu

David Čaněk (MemSource, david.canek@memsource.com)

Riešiteľ'ský kolektív

Dalibor Frívaldský (študent MFF UK, MemSource, dalibor.free@gmail.com)
študent MFF UK
študent MFF UK
študent MFF UK

Termín dokončenia

9 mesiacov od dátumu zahájenia projektu

Motivácia

Na trhu s CAT (Computer Assisted Translation, http://en.wikipedia.org/wiki/Computer-assisted_translation) nástrojmi prevládajú softwarové riešenia postavené na desktop platformách. Nevýhody takýchto aplikácií sú aj na strane používateľov aj na strane firiem, ktoré za nimi stoja. Používatelia sú nútení inštalovať dodatočný softvér na svoj počítač a starať sa o jeho údržbu a aktualizáciu. Z tohoto faktu vyplýva hlavná nevýhoda pre firmy implementujúce (nielen) CAT nástroje – softvér musia vyvíjať pre širokú paletu rôznych platforiem a zároveň poskytovať ľahký prístup ku novým verziám nástrojov. Proces zavádzania novej verzie však len zriedkavo pokryje celú bázu používateľov a tak je bežné, že je súčasne používaných niekoľko verzií tej istej aplikácie. Toto prináša komplikácie a zvýšené náklady pre údržbu a podporu. Riešenie poskytuje webová platforma, pomocou ktorej je možné odstrániť väčšinu problémov spojených s distribúciou CAT nástrojov zákazníkom.

Čo je to CAT editor prekladu

Proces prekladu textu zo *zdrojového* do *cieľového* jazyka za pomoci CAT nástrojov začína jeho rozdelením na krátke úseky textu – *segmenty*. Ich dĺžka je rôzna, no spravidla nepresahuje veľkosť jednej vety. Takto rozdelený text je prezentovaný prekladateľovi ako dvojstĺpcová *prekladová tabuľka*. Prvý (ľavý, zdrojový) stĺpec obsahuje *zdrojové segmenty* z rozdeleného textu. Druhý (pravý, cieľový) je určený pre preložený text reprezentujúci *zdrojový segment* v *cieľovom* jazyku (*cieľový segment*). Pri dopĺňaní prekladov má prekladateľ prístup ku najdôležitejšej časti CAT nástrojov – *prekladovej pamäti*. Tá obsahuje preklady *segmentov* z minulých dokumentov. Pri jednom preklade môže prekladateľ používať viacero *pamätí*.

CAT editor prekladu plní v tomto kontexte 2 hlavné úlohy:

- poskytuje pohľad na *prekladovú tabuľku* a umožňuje jej pohodlnú editáciu
- poskytuje kontextový prístup ku *prekladovým pamätiam*. To znamená, že prekladateľovi je zobrazovaná iba tá časť *pamäte*, ktorá súvisí s aktuálne prekladaným *segmentom*

Cieľ projektu

Cieľom projektu je vytvoriť webový editor prekladov, ktorý by svojou funkcionalitou odpovedal zaužívaným desktopovým riešeniam a odstraňoval ich nedostatky spojené s distribúciou. Editor by mal riešiť iba samotnú editáciu prekladovej tabuľky a prístup ku pamätiam. Všetky ostatné funkcionality spojené s prekladovým procesom (ako správu prekladov, segmentáciu, správu prekladových pamätí, ...) budú poskytovať webové služby firmy MemSource, s ktorými by mal editor spolupracovať. Na editor sú kladené nasledovné požiadavky:

- postavený na klient/server architektúre
 - klient by mal využívať, kde je to vhodné, moderné webové technológie (HTML5, CSS3, JavaScript, AJAX) a podporovať 2 najnovšie verzie najpoužívanejších webových prehliadačov (IE, Firefox, Chrome, Safari) na operačných systémoch Windows, GNU/Linux a Mac OS X
 - technológie použité na implementáciu servera musia byť kompatibilné s aspoň jedným z nasledujúcich poskytovateľov Cloud serverových riešení: Google App Engine, Amazon EC2
- podporovať aspoň 10 súčasne prebiehajúcich prekladov bez výrazného spomalenia
- podporovať plynulú editáciu prekladov o veľkosti aspoň 10 tisíc segmentov
- proces prekladania by nemal byť spomalený komunikáciou so serverom a celková práca s editorom by sa mala čo najviac priblížiť ku práci s desktopovým riešením. Jednotlivé operácie prebiehajúce nad prekladovou tabuľkou by mali byť čo najviac nezávislé na odpovediach zo serveru, aby nedochádzalo ku prerušeniu práce prekladateľa. Zároveň je potrebné docieľiť konzistentnosť stavu prekladu v prehliadači a na serveri. Vzhľadom na túto požiadavku je možné, ak sa to ukáže pri návrhu systému ako žiaduce, presúvať logiku medzi serverom a klientom. Z tohoto dôvodu nie je doleuvedené rozdelenie zodpovedností klienta/serveru pre projekt záväzná

Požiadavky kladené na server:

- spolupracovať so serverom pre správu prekladov pre prístup ku prekladom v bilingválnom formáte (XLIFF)
- transformovať preklady v hore uvedenom formáte do dátových štruktúr vhodných pre editáciu. Transformované dáta by mali mať nasledujúce vlastnosti:
 - poskytovať sekvenčný prístup ku segmentom v poradí, v akom sa nachádzajú v originálnom dokumente
 - umožňovať rozdeľovanie a zlučovanie segmentov
 - umožňovať filtrovanie segmentov na základe stavu alebo textového obsahu
 - udržiavať informácie o repetitions (repetition je zoznam segmentov, ktoré majú rovnaký zdrojový text) a umožniť propagáciu prekladu jedného segmentu na ostatné segmenty v repetition, ktoré sa v dokumente nachádzajú pod daným segmentom
 - texty v segmentoch majú svoju vlastnú štruktúru tvorenú pomocou párových a nepárových značiek. Tieto značky majú rôzny význam (napr. formátovanie časti textu na tučné písmo, kurzívu, podčiarknutie, indikátor spojenia segmentov, atď). Editor by mal tejto štruktúre rozumieť a viedieť s ňou efektívne narábať.
- udržiavať históriu undo/redo a umožňovať pohyb v nej pre každý prekladový proces. História by mala byť udržiavaná pre každú zmenu na úrovni jedného segmentu alebo skupiny segmentov v prípade, že sa jedná o hromadnú operáciu.
- umožniť rozšírenie o kolaboratívny preklad jedného dokumentu viacerými prekladateľmi (teda konkurentný prístup a editovanie prekladu z viacerých klientov). Plne funkčné kolaboratívne

editovanie je mimo rámec projektu

Požiadavky kladené na klienta:

- poskytovať pohodlný prístup ku prekladovej tabuľke daného dokumentu optimalizovaný pre webové prehliadače. Stránkovanie tabuľky prekladu nie je postačujúce pre pohodlný preklad. Jednotlivé časti tabuľky by sa mali dynamicky načítať zo serveru (a prípadne uvoľňovať z pamäti) podľa pohybu používateľa v preklade. Dôraz je kladený na čo najmenšie prerušenie práce načítaním dát zo serveru.
- nad prekladovou tabuľkou podporovať sadu operácií, ktoré možno vykonávať na jednom segmente alebo viacerých vyznačených segmentoch zároveň:
 - editovanie textu odpovedajúce základnej sade nástrojov pre prácu s textovým editorom s podporou formátovania, ktorý definujú párove formátovacie značky v segmente
 - zlučovanie segmentov
 - rozdeľovanie segmentov
 - skopírovanie zdrojového textu do cieľového
 - vymazanie cieľového textu
 - presun na najbližší nasledujúci segment alebo najbližší nasledujúci potvrdený alebo nepotvrdený segment (podľa preferencie prekladateľa)
 - potvrdenie segmentu. Preklad potvrdeného segmentu sa propaguje v repetition
 - vloženie značky zo zdrojového textu do cieľového
- poskytnúť efektívne rozhranie ku funkcionalite filtrovania na serveri editoru. Pre vyhľadávanie v texte je postačujúca funkcionalita webového prehliadača.
- poskytnúť prístup k undo/redo histórii zmien v preklade
- poskytovať kontextový prístup ku prekladovým pamätiam a strojovému prekladu pre aktuálne editovaný segment
- kontrolovať pravopis (v rozsahu, ktorý podporujú webové prehliadače, prípadne plugíny pre ne)
- podporovať jazyky písané zľava doprava aj sprava doľava

Webové služby firmy MemSource

Výsledný editor bude závislý na viacerých webových službách firmy MemSource. Aby bola zabezpečená nezávislosť úspešného dokončenia projektu od vývoja a existencie firmy MemSource, firma MemSource poskytne riešiteľskému tímu prístup ku zdrojovým kódom a buildom služieb vyžadovaných pre prácu editoru na vývojové, testovacie a prezentačné účely.