LetsMT!

Platform for Online Sharing of Training Data and Building User Tailored Machine Translation

prof. dr. sc. Marko Tadić
Sveučilište u Zagrebu
Filozofski fakultet
Odsjek za lingvistiku

ICT-PSP dani
Zagreb, 2011-03-14

Strojno prevođenje danas

- **statističko strojno prevođenje** (SMT) ostvarilo proboj u razvoju sustava za **strojno prevođenje** (MT)
  - npr. Google Translate, Microsoft (Bing) Translator...
- kvaliteta SMT sustava u mnogome ovisi o količini podataka za njihovo treniranje
  - podatci = paralelni tekstovi (izvornik i prijevod)
  - visokoovisni o domeni i tematici kojom se tekst bavi
- umanjivanje netočnosti SMT-sustava
  - većom količinom tekstova za treniranje
  - treniranjem na tekstovima iz uskospecijaliziranih domena
- dostupni paralelni tekstovi na www-u samo su manji dio ukupno postojećih paralelnih tekstova
  - smješteni lokalno u sustavima tvrtki, ustanova, pojedinaca...

Prilagodba SMT sustava?

- mrežno dostupni SMT sustavi
  - opće naravi tj. trenirani na općem jeziku
  - loše prevode specifične tekstove iz uskih domena
  - loše prevode pojedine jezične parove zbog premale količine podataka za treniranje
- prilagodba za specifične potrebe pojedinih korisnika
  - neisplativo preskupa za male tvrtke i većinu ustanova
- lokalizacijska industrija
  - ne može u cijelosti iskoristiti svoje paralelne resurse

Vlastiti sustavi za MT?

- open-source rješenja
  - Moses i Giza++
  - omogućuju “besplatnu” izgradnju vlastitih SMT sustava
- ali, još uvijek je potrebna
  - lokalna infrastruktura (ljudi i strojevi)
  - ekspertiza u izgradnji SMT sustava
  - ekspertiza u primjeni SMT sustava
  - neisplativo preskupo za male tvrtke i većinu ustanova

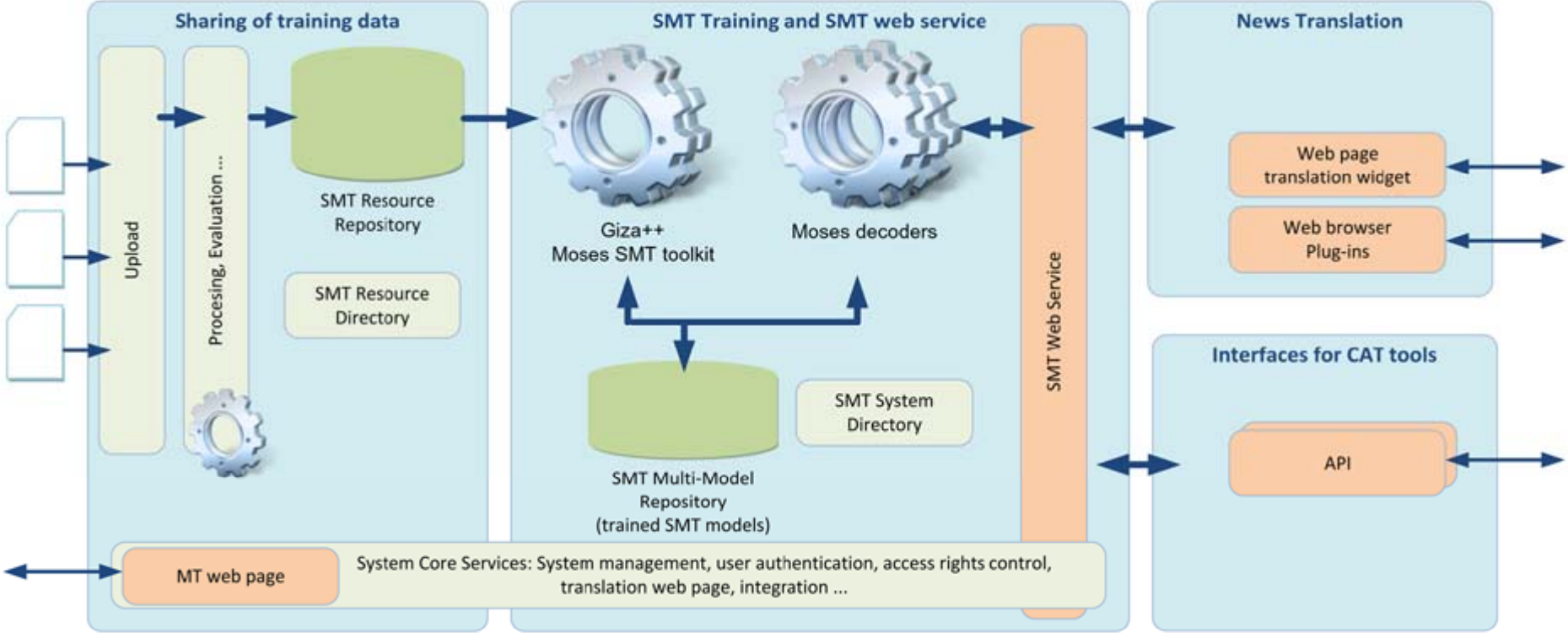
Vizija projekta LetsMT!

- izgraditi inovativnu suradnu **platformu za razmjenu podataka** (paralelnih tekstova) i **izgradnju sustava za SMT**
- platforma u *oblaku* služi za **generiranje višestrukih SMT-sustava** na temelju javnih ili zaštićenih tekstova
- očekujemo porast
  - kvalitete prijevoda
  - broja domena pokrivenih SMT-sustavima
  - broja jezika za koje postoje SMT-sustavi u uskospecijaliziranim domenama
- usredotočenost na “male” europske jezike
  - letonski, litavski, estonski, švedski, danski, hrvatski, češki, slovački, nizozemski...

Temeljni scenarij

- osnovni sudionici i koraci
  - registrirani korisnik (tvrtka, ustanova ili pojedinac)
  - šalje svoje podatke (paralelne tekstove) LetsMT! sustavu
  - podatci mogu bili javni ili zaštićeni
  - na LetsMT! platformi izgrađuje se statistički prijevodni model (T-model) na temelju poslanih tekstova i stvara se SMT-sustav
  - SMT-sustav izgrađen na javnim podatcima koriste
    - svi korisnici
  - SMT-sustav izgrađen na zaštićenim podatcima koristi
    - korisnik koji ga je stvorio
    - ostali korisnici kojima je on dopustio pristup

Temeljni scenarij



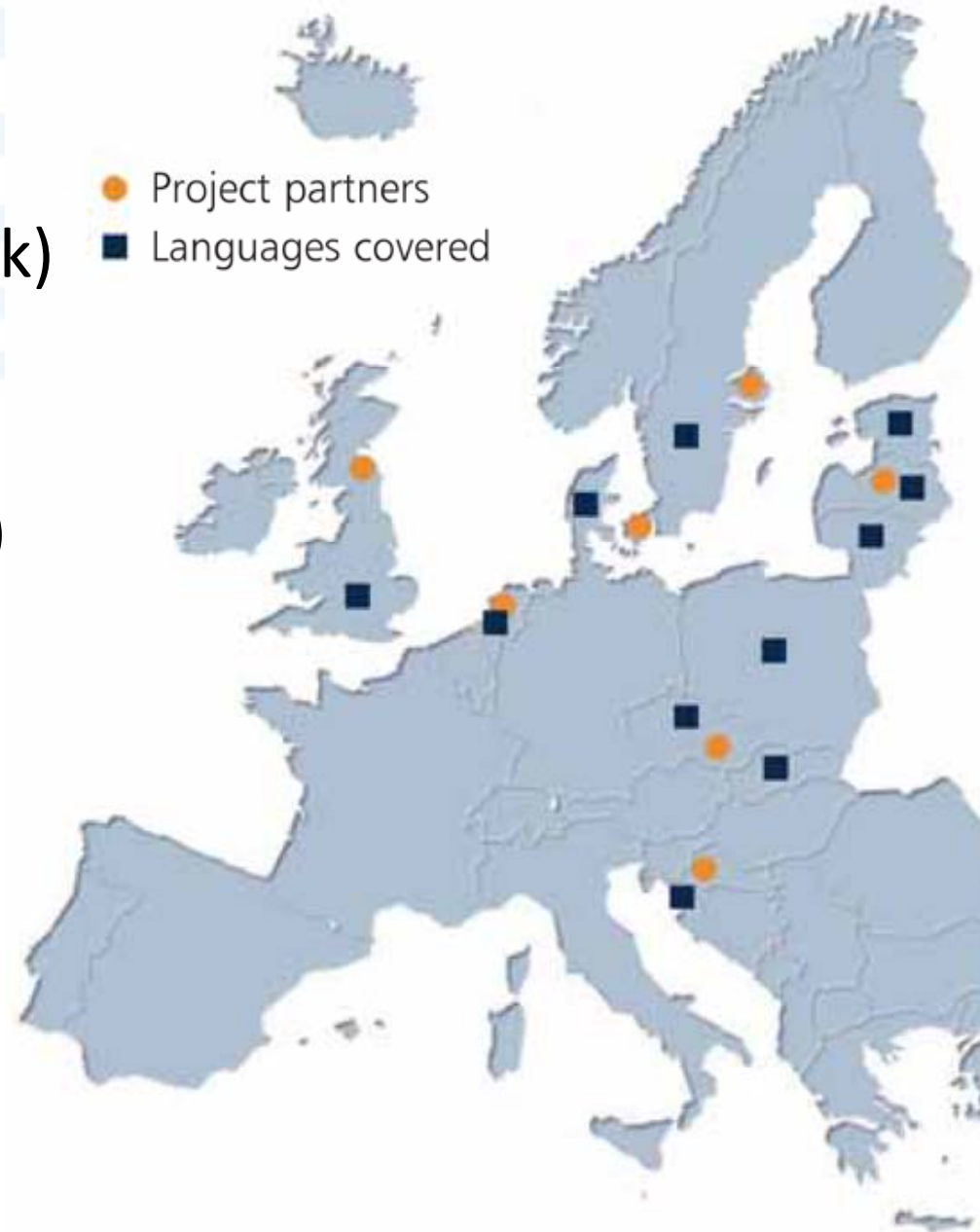
- *widget* za prijevod www-stranica
- priključak za *browsere* (Mozilla i IE 9.0)
- u perspektivi aplikacije za iOS i Android sustave

LetsMT! projekt

- područje
  - CIP-ICT-PSP.2009.5.1
  - Multilingual Web: Machine translation for the multilingual web
- ugovor br.
  - 250456
- trajanje
  - 2010-03-01 – 2012-08-31
- koordinator
  - Tilde, Riga
- ukupna vrijednost projekta
  - 3,340 M€, EK pokriva 1,670 M€
  - hr partner: 0,414 M€, EK pokriva 0,207 M€

LetsMT! partneri

- Tilde (lv)
- University of Edinburgh (uk)
- Sveučilište u Zagrebu, Filozofski fakultet (hr)
- Kopehagen University (dk)
- Uppsala University (se)
- Moravia (cz)
- SemLab (nl)



Problemi izvođenja projekata u RH

- neriješen status potpore EK u slučaju javnih ustanova
  - uposlenici (npr. na sveučilištu): 100% plaće iz proračuna
  - kolektivnim ugovorom određeno
    - smiju se u okviru 8-satnog radnog vremena baviti i istraživačkim radom
    - za to mogu biti dodatno nagrađeni
  - što sa sredstvima koja se *refundiraju* za rad tih uposlenika na projektima (sveučilište "iznajmljuje" uposlenike projektu)
    - vraćaju se u proračun?
    - raspodjeljuju se kao nagrada za suradnike na projektu?
- uprave i računovodstva ne rješavaju taj problem
  - što iz neznanja, što iz straha
  - blokada odvijanja projekata, neispunjenje ugovornih obveza
  - prijetnja povrata sredstava EK-u
  - negativna slika o hr ustanovama, RH u cjelini, novi projekti?

Problemi izvođenja projekata u RH

- hitno tražiti službeno mišljenje Ministarstva financija
  - o stavljanju na raspolaganje tih sredstava istraživačkim timovima
- očekujemo u tome svu moguću pomoć Središnjega državnoga ureda za e-Hrvatsku

ZAHVALJUJEM NA POZORNOSTI

The research within the project LetsMT! leading to these results has received funding from the **ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web**, grant agreement nº 250456.