Machine Translation of Medical Text in the KConnect Project

Petra Galuščáková, Jan Hajič, Jindřich Libovický,
Pavel Pecina, Aleš Tamchyna

Charles University in Prague
Institute of Formal and Applied Linguistics

KCONNECT

Introduction

- KConnect is a follow-up project of Khresmoi
- goals: provide components developed in Khresmoi as commercialized cloud services
- role of MT: provide cross-lingual search and access to medical documents
  - search queries
  - document summaries

Training Data

- new languages:
  - Swedish, Spanish, Polish, Hungarian
- in-domain corpora collected and processed
  - UMLS, EMEA, MuchMore, Wikipedia, PatTR, COPPA, Mesh, subtitles,...

Training Data: Statistics

	parallel		monolingual only	
	in-domain	general domain	in-domain	general domain
cs	21	665	1	93
de	126	310	4	699
es	74	1248	2	474
fr	193	896	2	589
hu	19	641	1	98
pl	17	606	1	205
sv	24	409	21	158
en	–	–	6087	2100

Training data sizes, all figures are in millions of words.

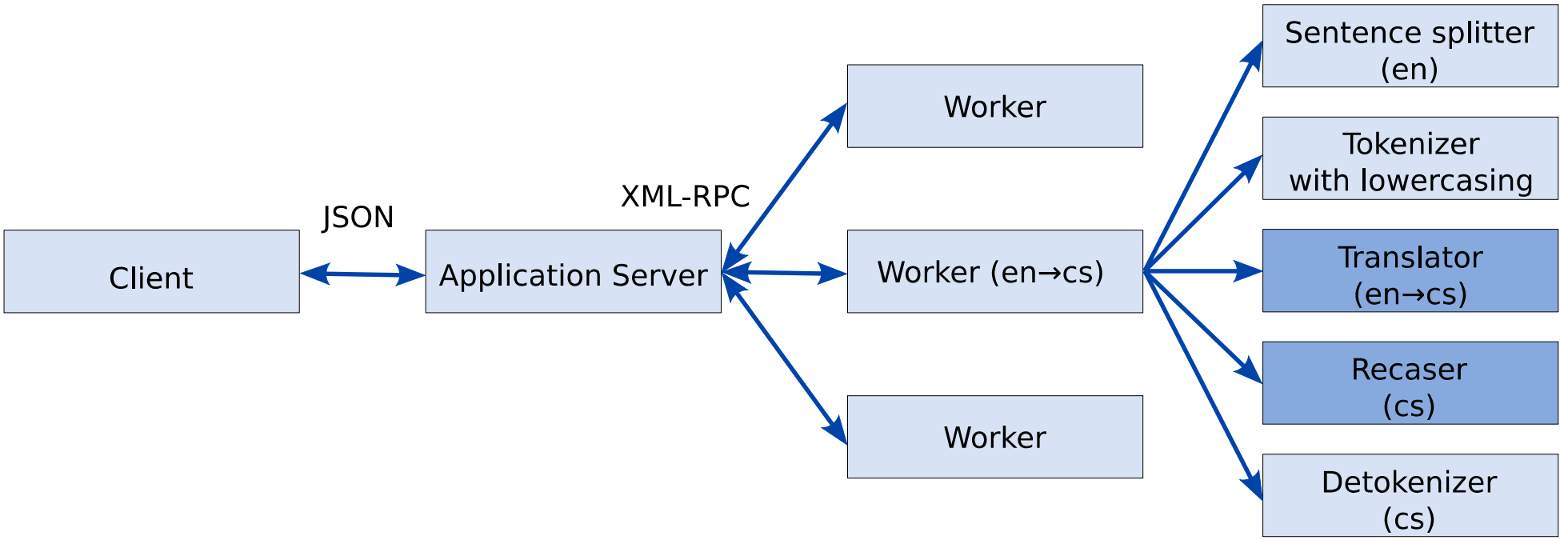
Domain Adaptation

- Data selection
  - divide data into „medical-like“ and „general“ parts

    (based on language model perplexity)
- Model interpolation
  - build separate models (phrase table, language model) for each part
  - use linear interpolation to combine them
    - SRILM
    - TMCombine

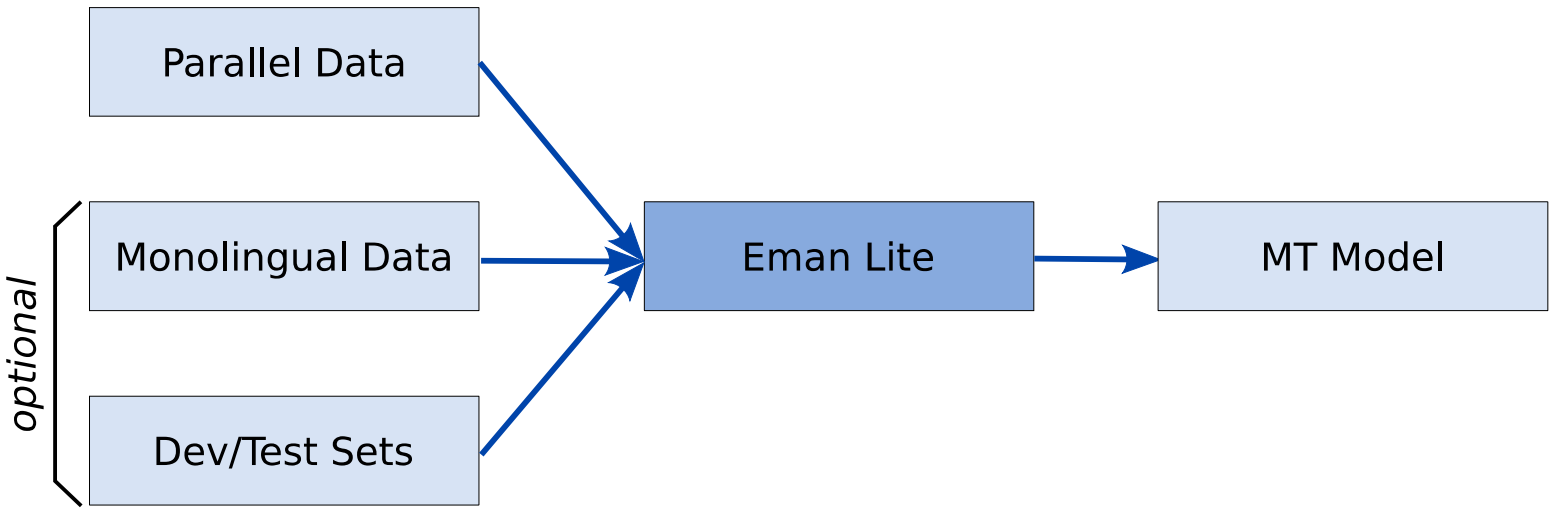
MT as a Web Service

- MTMonkey
- developed within Khresmoi, now actively extended and maintained
- runs in a cluster of 20 servers



Training Toolkit

- Eman Lite
- fully automated MT system training
- command-line application implemented
- goal: web-based interface, tight integration with MTMonkey



Thank you!

Questions?