Achim Ruopp
Digital Silk Road LLC
achim@digitalsilkroad.net

BIGTM.NET
GAINING A HEAD START ON TRANSLATION PROJECTS

A New Translation Project

- "The Bundle Breaker® is a sturdy machine with moving parts driven by electric motors, pneumatics, and hydraulics. These parts move automatically when the machine is turned on."

- "The downstream platen of the compression bridge may move side-to-side unexpectedly and strike personnel in its path."

A New Translation Project
Bi-Lingual Terminology

- Criteria
  - In domain
  - Correct
  - Current
  - In context
- Sources
  - Translation memory from previous projects
  - Domain dictionaries
  - … BigTM.net

BigTM.net
Custom Translation Search

- What is it?
  - Custom Translation Search Engine
  - Input: project source text
  - Searches the web for similar bi-lingual pages
  - Indexes discovered bi-lingual pages
  - Provides search UI
- Current Language Pairs
  - English - French / Italian / German / Spanish

BigTM.net Project Page

BigTM.net Search Results

Terminology Criteria

- In domain
  - BigTM.net searches the web for similar parallel pages
  - Search UI results are ranked according to similarity with source text
- Correct
  - Translator can choose from discovered variants
  - Relies on web data
- Current
  - Web search turns up new translations
- In context
  - Search results always displayed in context

BigTM.net Architecture

BigTM.net Data Flow

Web Search with Extracted Terms

Page Pair Candidates

Matcher

Parallel Corpus

Classifier

Glossary/ TM

Aligner

Searchabl e Index

MT System

CAT Tool
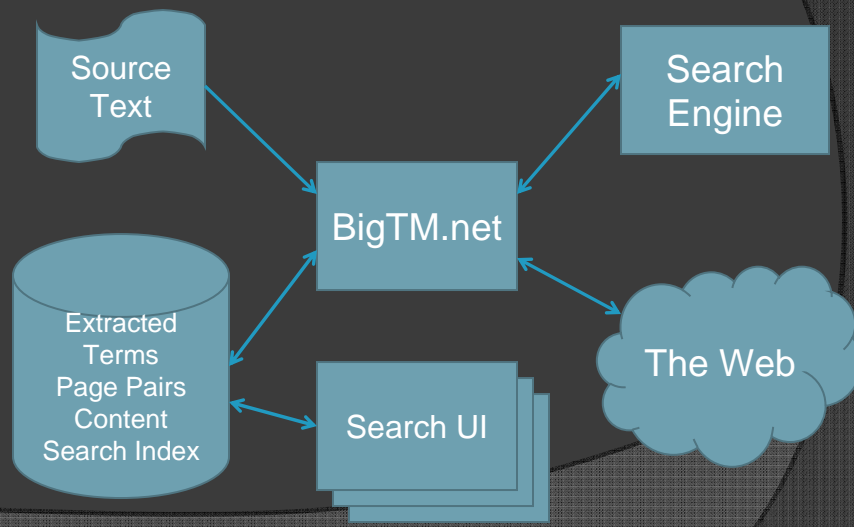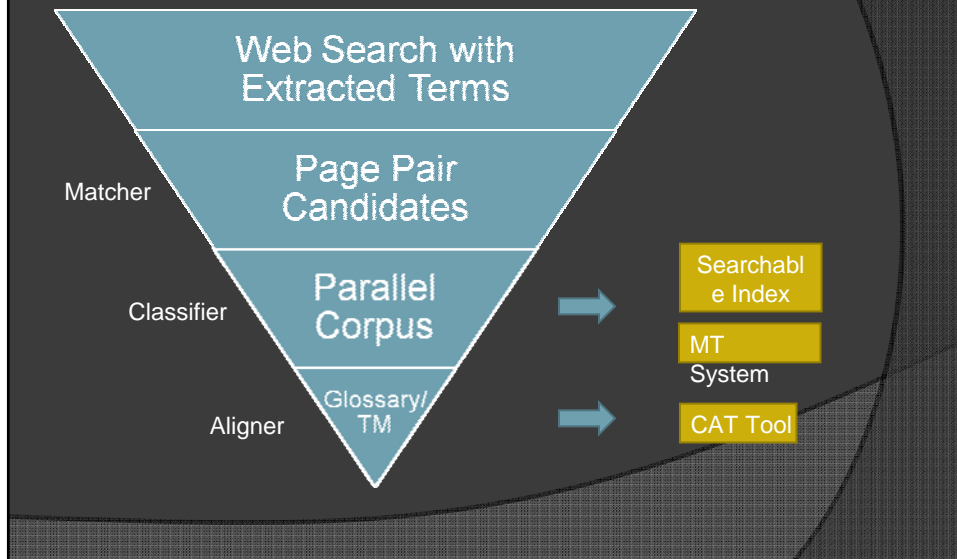
Similarity with Source Text

Integration with Translation Tools

- Tools with web search integration
- Dictionary download
  - CSV format with probabilities
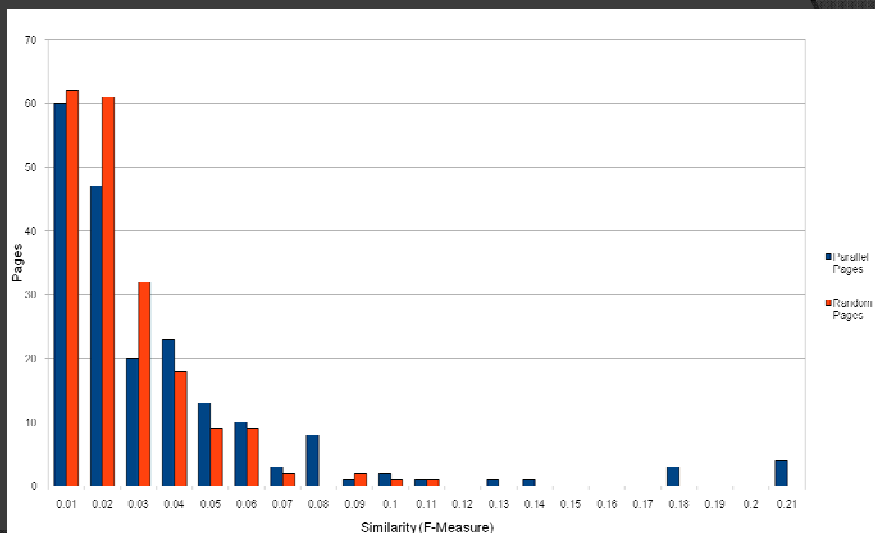- Translation memory download
  - Standard TMX format
  - Best used with sub-segment matching

Training of domain-specific statistical MT systems

- Supplement general domain corpus with domain-specific corpus downloaded from BigTM.net
- KDE documentation prototype English-German

Corpus	MT System	BLEU
~ 600k sentences Europarl	Moses	12.81
~ 400k sentences Europarl ~ 200k sentences from BigTM	Moses	16.47 (+3.66)
n/a	Google	18.29 (+5.48)

Benefits Summary

- Provides a searchable index for translations of similar texts
- Rapidly prototypes terminology for review by the customer
- Provides a core translation memory as a starting point
- Allows the training of domain-specific machine translation systems

References

- Website: http://www.bigtm.net/
- Email: bigtm@digitalsilkroad.net
- Email your source text to get added to the beta program