High quality language solutions delivered on time,
. . . with a smile!

ALS Language Technologies for User-Generated Content

Sergio Penkale
Applied Language Solutions

Outline

Outline

Why MT for User-Generated Content?

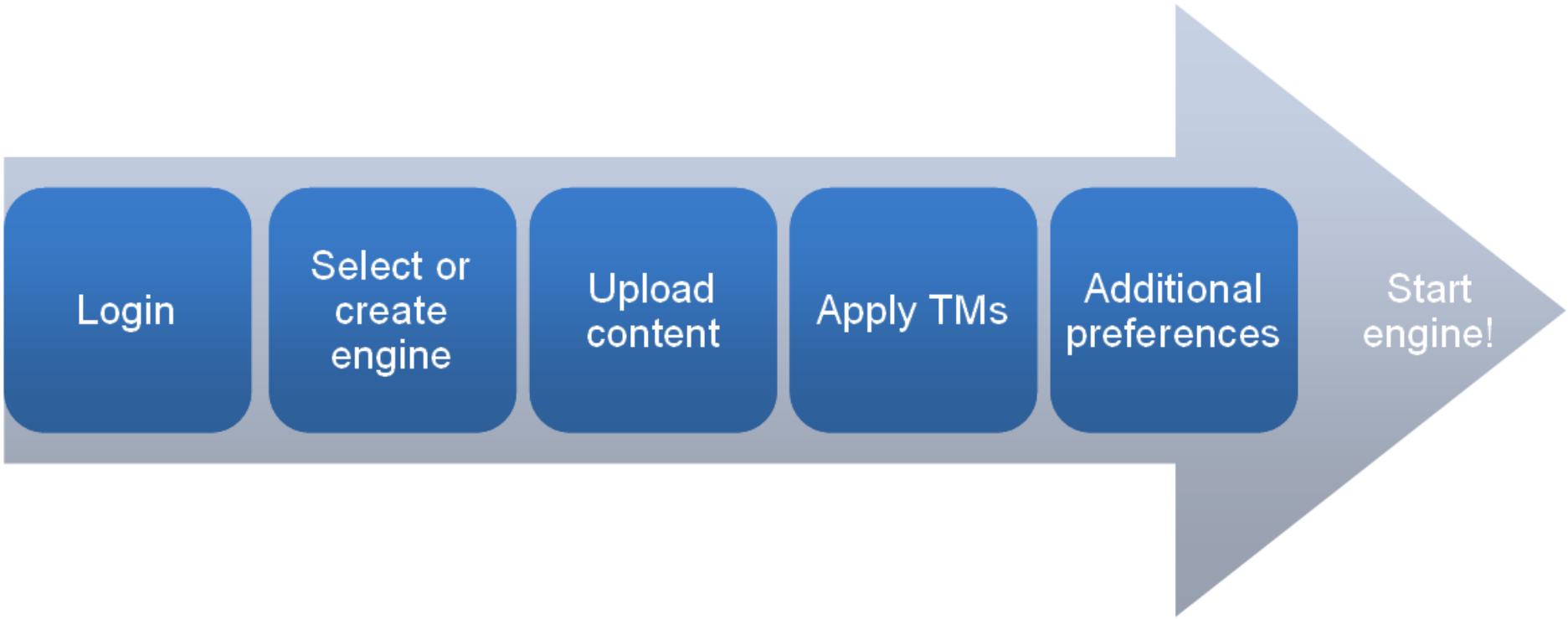
- Web 2.0 is UGC-centric
- English-language users only 27% of Web population
- Currently most of this content remains untranslated
- MT of UGC enables new markets penetration

Challenges

- Translating UGC is problematic
- Much of it is of low quality
  - Non-native speakers (grammar, morphology)
  - Native speakers, involuntary errors (typos)
  - Native speakers, deliberate errors (e.g. LOL, c@@l)
- Typically parallel data not available

Outline

SmartMATE: self-serve translation platform



SmartMATE: Components

- Moses-powered MT
  - Encoding conversion
  - Data cleaning
  - Markup handling
  - Handling of special entities (URLs, special characters, etc.)
  - Tokenisation
  - Lowercasing/Truecasing
  - Glossary Injection
  - Moses training in the cloud
  - Moses decoding in the cloud
- TMs
- File Filtering
- Terminology
- Web UI with online editor
- API

Outline

Translation of Social Games

- Work for a large online games developer and publisher
- ALS handles TM, MT, and Translation. QA is performed by a third party
- Localized into 15 lang. pairs for more than 180 million users
- Fully deployed within SmartMATE
- Early stages: still no MT

Monolingual MT

- Games written in English by non-native developers
- English is of bad quality
- “bad” English into “good” English translation required before source can be localized
- Large collection of corrected sentences $\Rightarrow$ MT training data!

Outline

Case Study 2

- Work facilitating multilingual solution for a large social network provider
- Users can communicate with each other in the available languages
- So far, MT implemented between: English, Russian, Arabic, Turkish

Data Collection

- Client's in-domain monolingual data for Language Models (LMs)
- OPUS subtitle data as parallel corpus
- Development and test sets created based on LM perplexity
- Twitter data as in-domain LM
- Manual translation of slang dictionaries (over 5K entries)

Data Collection

- Client's in-domain monolingual data for Language Models (LMs)
- OPUS subtitle data as parallel corpus
- Development and test sets created based on LM perplexity
- Twitter data as in-domain LM
- Manual translation of slang dictionaries (over 5K entries)

Data Collection

- Client's in-domain monolingual data for Language Models (LMs)
- OPUS subtitle data as parallel corpus
- Development and test sets created based on LM perplexity
- Twitter data as in-domain LM
- Manual translation of slang dictionaries (over 5K entries)

Data Collection

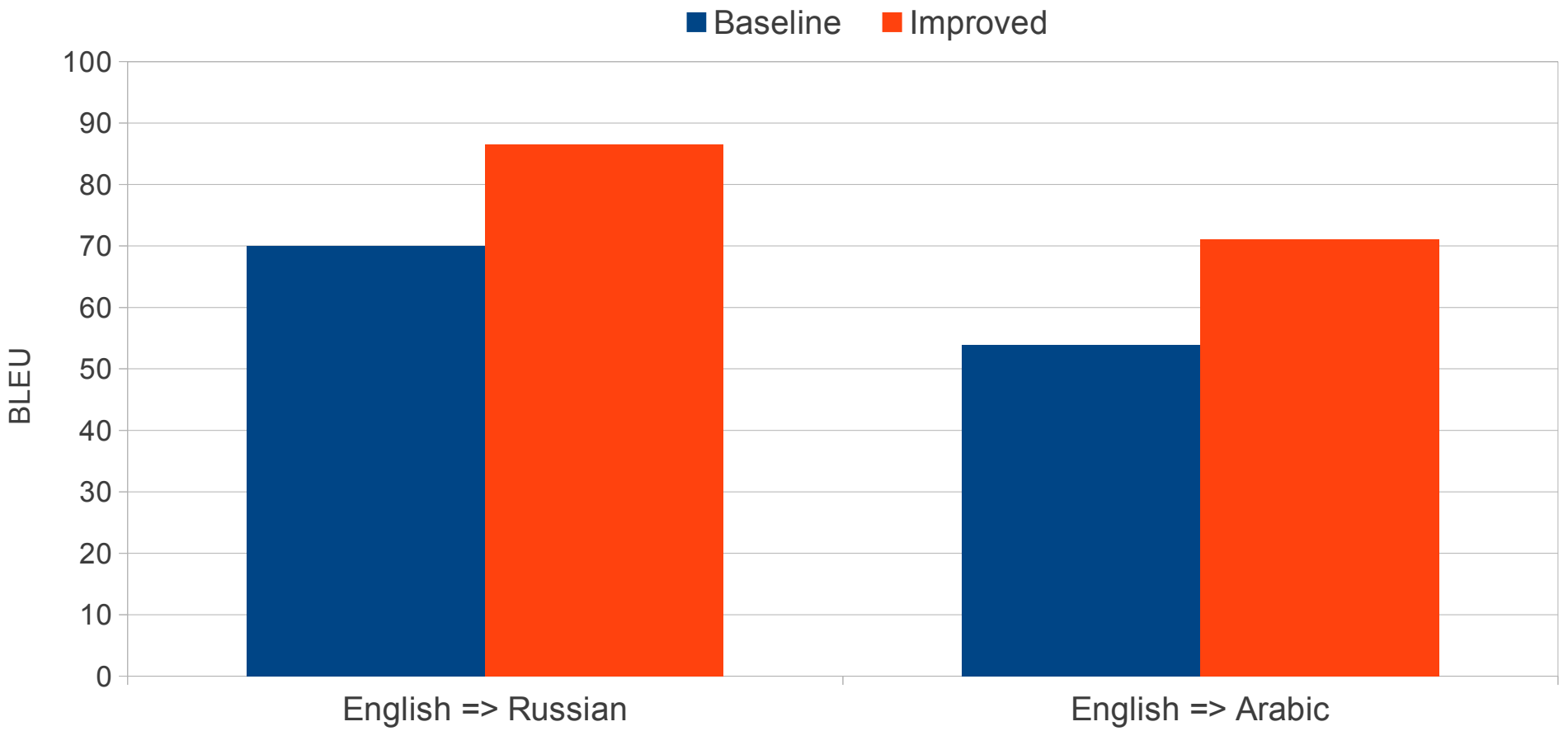
- Client's in-domain monolingual data for Language Models (LMs)
- OPUS subtitle data as parallel corpus
- Development and test sets created based on LM perplexity
- Twitter data as in-domain LM
- Manual translation of slang dictionaries (over 5K entries)

Data Pre-processing and Cleaning

- Usual Cleaning Steps
  - Correct character encoding
  - Deal with formatting tags
  - Remove duplicate sentences
  - Check Source:Target length ratio
  - Generalizing URLs, emails, etc
- Spellchecker (edit distance)
- Soundex Algorithm

Misspelled	Soundex Code	Correct Spelling	Soundex Code
c@@l	C400	cool	C400
tmrw	T560	tomorrow	T560
whatdoyouwant	W312	what do you want	W312

Development Results



Final Results

Language Pairs	BLEU ($\Rightarrow$)	BLEU ($\Leftarrow$)
English $\Leftrightarrow$ Russian	86.49	91.01
English $\Leftrightarrow$ Arabic	71.10	88.39
English $\Leftrightarrow$ Turkish	79.65	80.78
Arabic $\Leftrightarrow$ Russian	78.29	72.30
Arabic $\Leftrightarrow$ Turkish	73.07	68.06
Russian $\Leftrightarrow$ Turkish	90.54	88.72

Incorporating Feedback

- Client provided suggestions and some post-edited parallel data
- Most feedback about lexical choice. E.g. **“nice”** vs. **“Nice” (France)**

- Corpus editing and engine retraining
  - Regular expressions created
  - Added post-edited data to modified corpus

Usage Statistics

- "always on" online translation
- Statistical prunning (Johnson et. al., 2007) to meet speed demands
- Client Connects to SmartMATE through REST API
- Translated 135 Million words in 7 months

Time	Translated Words
02/2012	71,779
03/2012	16,182,075
04/2012	16,608,694
05/2012	23,298,287
06/2012	18,843,487
07/2012	36,952,204
08/2012	23,301,706

Outline

Summary

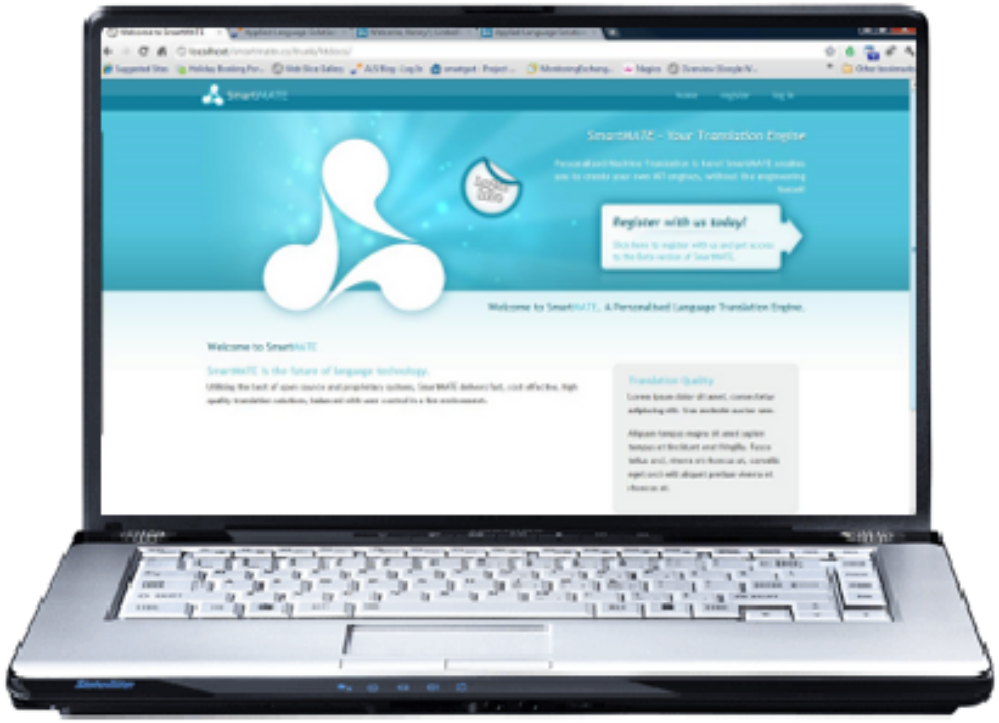
- Presented overview of ALS capabilities for UGC
- Presented use cases of MT for UGC
- Engines for social media provider currently live and being heavily used for 12 language pairs
- Domain adaptation/data crawling and pre-processing/cleaning enable high-quality MT for UGC

Future Work

- Improve handling of morphology
- Implement regular expressions for wordplay (e.g. "**coool**")
- Named Entity handling
- Do-Not-Translate foreign expressions (e.g. "**al dente**")

Applied Language SOLUTIONS

30-day free trial



www.smartmate.co

Applied Language SOLUTIONS

Thank you for your attention!