Improving SMT for Baltic Languages with Factored Models

Raivis SKADIŅŠ[a,b], Kārlis GOBA[a] and Valters ŠICS [a]
[a] *Tilde SIA, Latvia*
[b] *University of Latvia, Latvia*

The Fourth International Conference
HUMAN LANGUAGE TECHNOLOGIES — THE BALTIC PERSPECTIVE
Riga, Latvia, October 7–8, 2010

In this presentation

- Current situation with Latvian & Lithuanian MT
- Motivation of this research
- SMT with factored models
  - English-Latvian
  - Lithuanian-English
- Evaluation
- The latest improvements

Current situation

- Latvian
  - MT in Tildes Birojs 2008 (RBMT)
  - Google Translator (SMT)
  - Microsoft Translator (SMT)
  - Pragma (RBMT)
  - IMCS system (SMT)
- Lithuanian
  - Google Translator (SMT)
  - Bing Translator (SMT)
  - VMU system (RBMT)

Motivation

- Both Latvian and Lithuanian
  - Morphologically rich languages
  - Relatively free order of constituents in a sentence
- Small amount of parallel corpora available
- We were not happy with a quality of existing MT

- Goal
  - not to build yet another SMT system using publicly available parallel corpora and tools
  - to add language specific knowledge to assess the possible improvement of translation quality

SMT is not so easy

- There are good open source tools (Giza++, Moses etc.) and even some training data available (DGT-TM, OPUS)
- Why it is not so easy to build SMT for Baltic languages
  - Rich morphology
  - Limited amount of training data
- Translating from English
  - How to chose the right inflected form
  - How to ensure agreement
  - How to deal with long distance reordering
- Translating to English
  - Out of vocabulary issue
  - How to deal with long distance reordering

Translating from English

English-Latvian

- The main challenge – inflected forms and agreement
- Simple SMT methods relay on size of training data
- Factored methods allow integration of language specifics
  - Lemmas, morphology, syntactic features, …
- There is no one best way how to use factored methods
- Solution depends on language pair and available tools

Experiment 1: English-Latvian

- Training data:

Bilingual corpus	Parallel units
Localization TM	~1.29 mil.
DGT-TM	~1.06 mil.
OPUS EMEA	~0.97 mil.
Fiction	~0.66 mil.
Dictionary data	~0.51 mil.
Total	**4.49 mil.**
	(3.23 mil. filtered)

Monolingual corpus	Words
Latvian side of parallel corpus	60M
News (web)	250M
Fiction	9M
Total, Latvian	**319M**

Experiment 1: English-Latvian

- Development and evaluation data
  - Development - 1000 sentences
  - Evaluation – 500 sentences
  - Balanced

Topic	Percentage
General information about European Union	12%
Specifications, instructions and manuals	12%
Popular scientific and educational	12%
Official and legal documents	12%
News and magazine articles	24%
Information technology	18%
Letters	5%
Fiction	5%

- Tools
  - GIZA++, Moses, SRILM
  - Latvian morphological tagger developed by Tilde

Experiment 1: English-Latvian

- Factored models
  - More than 10 different models tried
  - Here presented (1) gives good results and (2) is reasonably fast

System	Translation Models	Language Models
EN-LV SMT baseline	1: Surface → Surface	1: Surface form
EN-LV SMT suffix	1: Surface → Surface, suffix	1: Surface form 2: Suffix
EN-LV SMT tag	1: Surface → Surface, morphology tag	1: Surface form 2: Morphology tag

Experiment 1: English-Latvian

- Automatic evaluation

System	Language pair	BLEU
Tilde rule-based MT	English-Latvian	8.1%
Google	English-Latvian	32.9%
Pragma	English-Latvian	5.3%
SMT baseline	English-Latvian	24.8%
SMT suffix	English-Latvian	25.3%
SMT tag	English-Latvian	25.6%

- Human evaluation

System1	System2	Language pair	p	ci
SMT tag	SMT baseline	English-Latvian	58.67 %	±4.98 %
Google	SMT tag	English-Latvian	55.73 %	±6.01 %

Translating to English

Lithuanian-English

- The main challenge – out of vocabulary
- Simple SMT methods relay on size of training data

- We do not have a morphologic tagger for Lithuanian
- Simplified approach – splitting each token into two separate tokens containing the stem and an optional suffix.
- The stems and suffixes were treated in the same way in the training process.
- Suffixes were marked to avoid overlapping with stems.

Experiment 2: Lithuanian-English

- Training data:

Bilingual corpus	Parallel units
Localization TM	~1.56 mil.
DGT-TM	~0.99 mil.
OPUS EMEA	~0.84 mil.
Dictionary data	~0.38 mil.
OPUS KDE4	~0.05 mil.
Total	**3.82 mil.**
	(2.71 mil. filtered)

Monolingual corpus	Words
English side of parallel corpus	60M
News (WMT09)	440M
LCC	21M
Total, English	**521M**

Experiment 2: Lithuanian-English

- Development and evaluation data
  - Development - 1000 sentences
  - Evaluation – 500 sentences
  - Balanced (the same set of English sentences as before)

Topic	Percentage
General information about European Union	12%
Specifications, instructions and manuals	12%
Popular scientific and educational	12%
Official and legal documents	12%
News and magazine articles	24%
Information technology	18%
Letters	5%
Fiction	5%

- Tools
  - GIZA++, Moses, SRILM
  - A Simple Lithuanian stemmer developed by Tilde

Experiment 2: Lithuanian-English

- Models

System	Translation Models	Language Models
LT-EN SMT baseline	1: Surface → Surface	1: Surface form
LT-EN SMT Stem/suffix	1: Stem/suffix → Surface	1: Surface form
LT-EN SMT Stem	1: Stem → Surface	1: Surface form

Experiment 2: Lithuanian-English

- Automatic evaluation

System	Language pair	BLEU
Google	Lithuanian-English	29.5%
SMT baseline	Lithuanian-English	28.3%
SMT stem/suffix	Lithuanian-English	28.0%

System	Language pair	OOV, Words	OOV, Sentences
SMT baseline	Lithuanian-English	3.31%	39.8%
SMT stem/suffix	Lithuanian-English	2.17%	27.3%

- Human evaluation

System1	System2	Language pair	p	ci
SMT stem/suffix	SMT baseline	Lithuanian-English	52.32 %	±4.14 %

Conclusions

- Translating from English
  - Human evaluation shows a clear preference for factored SMT over the baseline SMT
  - However, automated metric scores show only slight improvement
- Translating to English
  - Simple stem/suffix model helps to reduce number of untranslated words.
  - The BLEU score slightly decreased (BLEU 28.0% vs 28.3%)
  - OOV rate differs significantly.
  - Human evaluation results suggest that users prefer lower OOV rate despite slight reduction in overall translation quality in terms of BLEU score.

Latest results

- English-Latvian and Latvian-English systems have been released: http://translate.tilde.com

- BLEU scores

System	Language pair	BLEU
translate.tilde.com	English-Latvian	33%
translate.tilde.com	Latvian- English	41%

- Human evaluation

System1	System2	Language pair	p	ci
Google	translate.tilde.com	Latvian-English	56.73 %	±4.60 %
Google	translate.tilde.com	English-Latvian	51.16 %	±3.62 %