Experiments with Domain Adaptation Methods for Statistical MT: From European Parliament Proceedings to Finnish Newspaper Text

Marcus Dobrinkat, Jaakko Väyrynen

marcus.dobrinkat@tkk.fi, jaakko.j.vayrynen@tkk.fi

August 20, 2010

Aalto University
School of Science
and Technology

Introduction

Problem

- Often a lack of data for the target domain
- The distribution of the test data is related, but different from the training data
- Language differs in many ways (style, lexical choice, textual organization, etc.)

$\Rightarrow$ A problem that affects most modern NLP systems: performance drop for a new domain/genre of language

Research Question

- **How can out-of-domain and in-domain training data be combined most efficiently and effectively**, to improve the performance of a statistical machine translation system on in-domain test data?

Bilingual Corpus Data

Europarl Corpus v2

- based on European Parliament proceedings from 1996-2003
- 11 European languages, 20 million words per language
- we use the English-Finnish data,
  1.3 million aligned bilingual sentences

language	type	sentences	word tokens	word types	characters in million	type-token ratio
Finnish	raw	1 262 914	18 837 151	479 779	146	0.0255
English	raw	1 262 914	26 073 619	83 496	143	0.0032
Finnish	pp	865 732	17 183 927	455 359	133	0.0265
English	pp	865 732	23 863 424	78 944	131	0.0033

- long sentence removal affected 400k sentences pairs

Bilingual Corpus Data

istuntokauden uudelleenavaaminen	resumption of the session
julistan perjantaina joulukuun 17. päivänä keskeytetyn euroopan parlamentin istunnon avatuksi ja esitän vielä kerran vilpittömän toiveeni siitä , että teillä olisi ollut oikein mukava joululoma .	i declare resumed the session of the european parliament adjourned on friday 17 december 1999 , and i would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period .
kuten olette varmaan saattaneet huomata , vuodenvaihteeseen 2000 povattuja suuria tietokoneongelmia ei ilmennytkään . sen sijaan todella kauheat luonnonkatastrofit koettelivat kansalaisia joissakin unionimme maissa .	although , as you will have seen , the dreaded ' millennium bug ' failed to materialise , still the people in a number of countries suffered a series of natural disasters that truly were dreadful .
te olette esittäneet toiveen , että tästä asiasta keskusteltaisiin lähipäivinä tämän istuntojakson aikana .	you have requested a debate on this subject in the course of the next few days , during this part-session .

Example paragraph in the Europarl corpus. On the left side are the Finnish sentences, which are aligned with the English sentences on the right side.

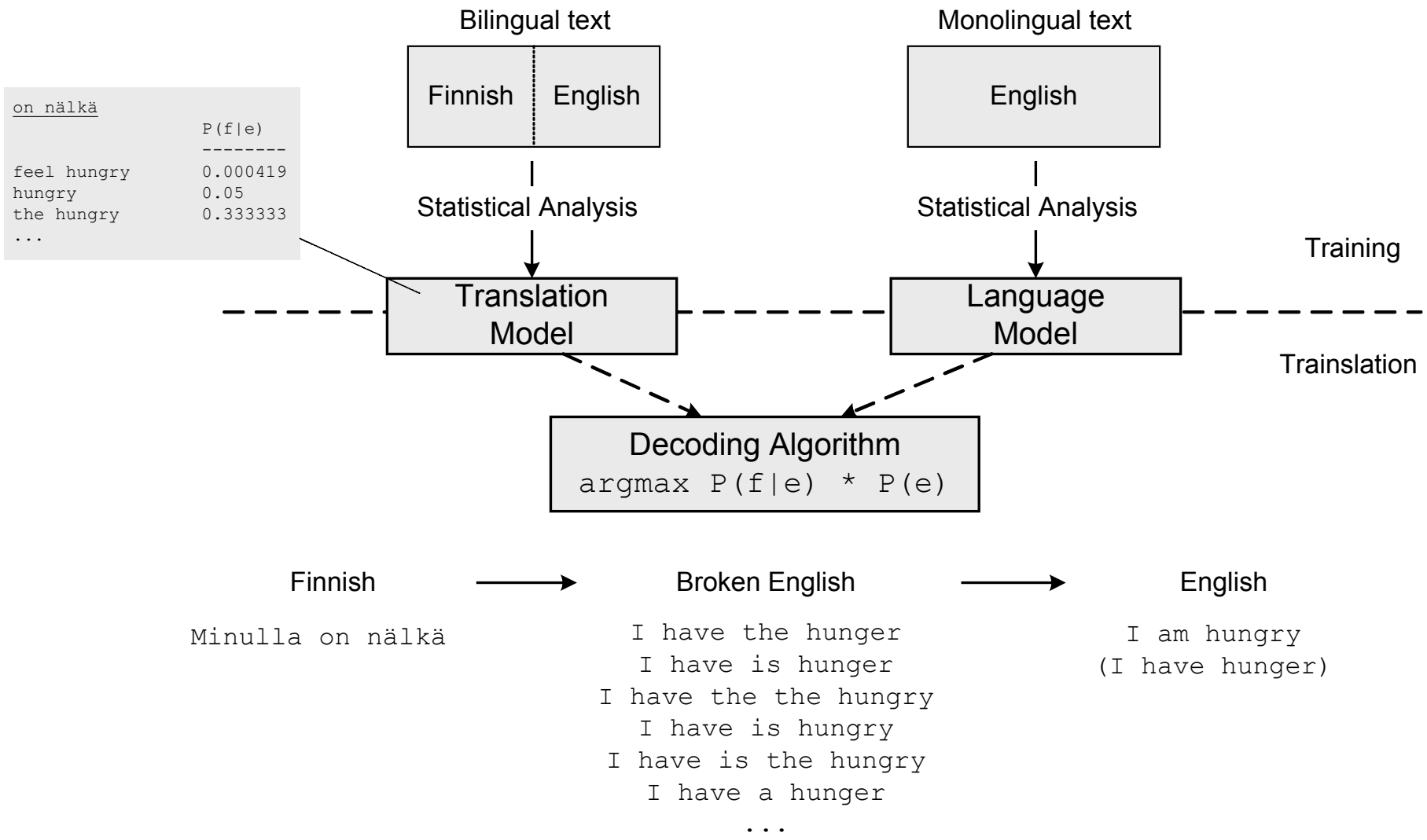
In-domain data

- News text domain (Iltalehti), 72k sentences
- Limited to short sentences (3..12 words restriction for easier human evaluation)
- Collected by the authors
- Parallel corpus has 1076 sentences

Europarl	Iltalehti
European commission typical: europaan, puhemies, komission komissio, parlamentin, unionin, unioni, neuvoston, parlamentti, jäsenvaltioiden, esittelijä, mietintö, mietinnössä	***Finnish given names:*** mika, kari, juha, pekka, jari, matti, jukka, paavo, antti, janne, mikko, lola, ari
More likely in longer sentences: että, jotka, täme, tämän, ja, jotta, tätä, tässä, sen, tästä, nämä	***Typical news words:*** poliisi, mm , elokuva, tv, ollut, tuli
	Colloquial language: mä

Distinctive word categories after a domain comparison between Europarl and Iltalehti corpora using log-likelihood ratio comparison.

Statistical Machine Translation Schema



Phrase-based Statistical Machine Translation Process

- Phrases achieve better local reorderings than word based models
- Log-linear models popular (based on a maximum entropy framework)

$$P(t|s) \propto \exp\left[\sum_{m=1}^{M} \lambda_m \, h_m(t,s)\right] \tag{1}$$

- Allow easy integration of arbitrary additional features

  $\Rightarrow$ typically a small number of often generative submodels used

Typical features used in an SMT system (e.g. Moses toolkit www.statmt.org)

- A phrase translation model $P(t|s)$ and the reverse $P(s|t)$
- A language model $P(t)$
- A word penalty model $P(t)$
- A reordering model $P(s,t)$

Adaptation Methods

How can MT performance be improved for a particular domain/genre?

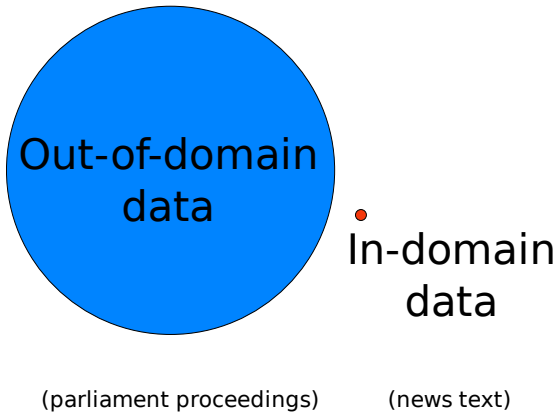
Obtain domain specific resources

- filter existing parallel corpora
- use domain terminology dictionaries
- take advantage of monolingual corpora
- collect feedback on-line

Combine domain specific resources

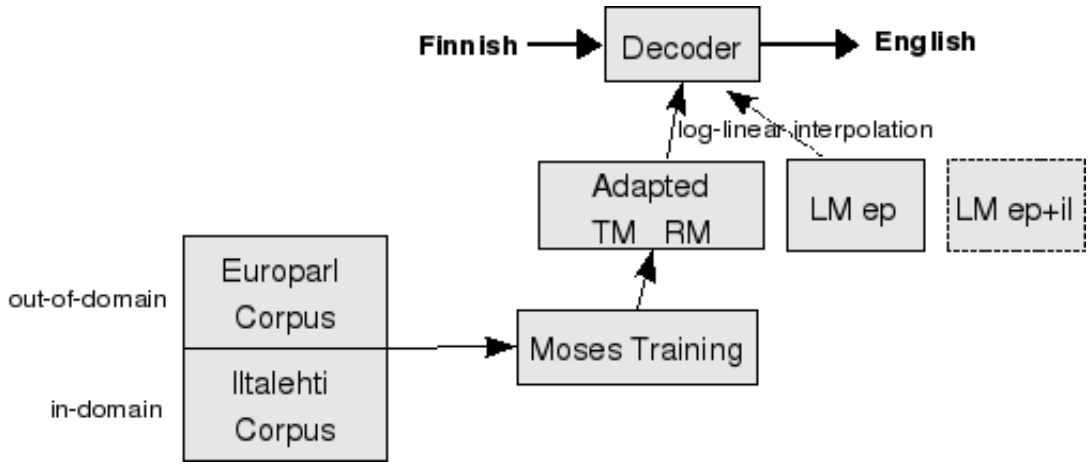
- combine data
- interpolate models
- post-edit automatically

$\Rightarrow$ our experiments compare the performance of these methods



Language Model Adaptation

LM adaptation only modifieds the language model of the system.

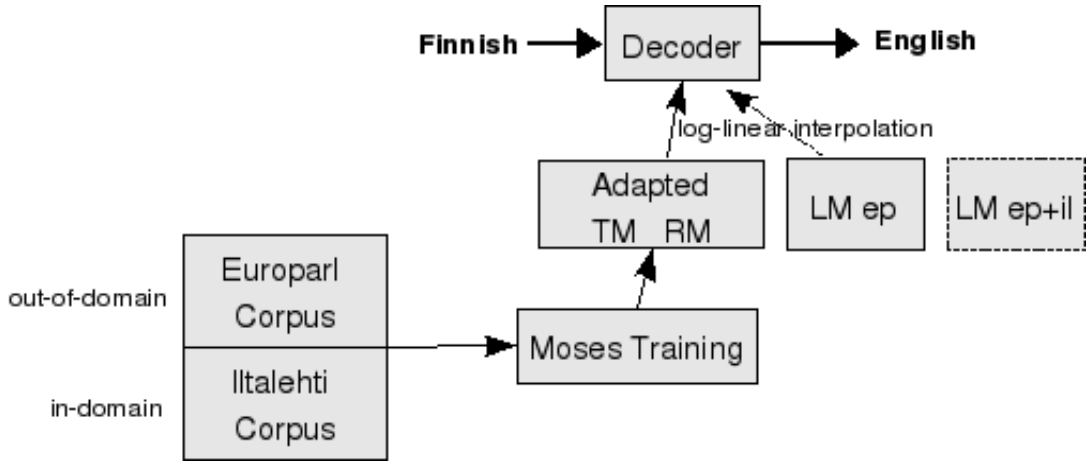


Experiments

- **L1** log-linear LM combination
- **L2** combined corpus LM
- **L3** linear LM interpolation

Translation Model Adaptation: **data combination**

The additional in-domain Iltalehti training data is added to the large baseline Europarl data.

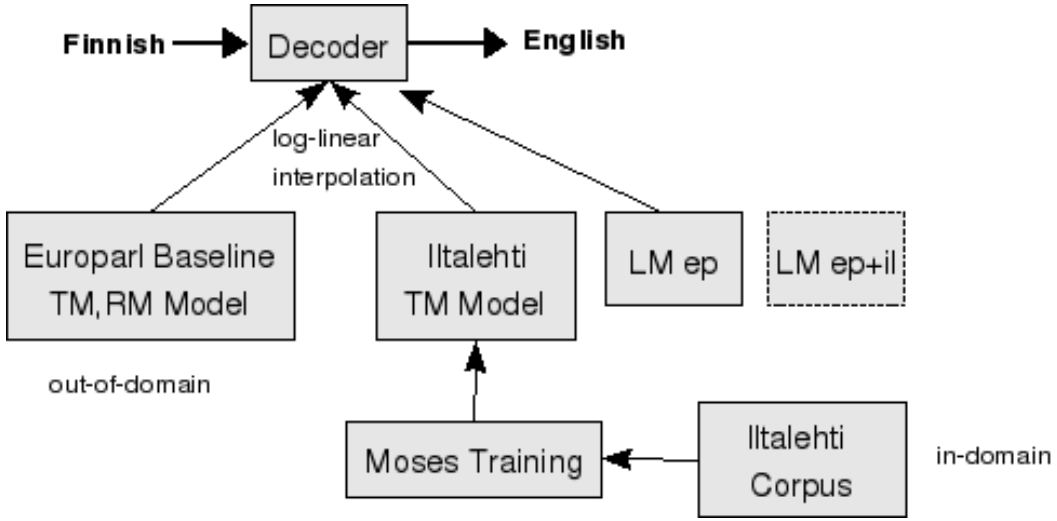


Experiments

- **C1** combined corpus translation model (TM) and reordering model (RM)
- **C2** C1 + combined corpus LM
- **C3** C1 + linear LM interpolation

Translation Model Adaptation: **model interpolation**

Translation and Reordering models are used as separate features in the log-linear model.



Experiments

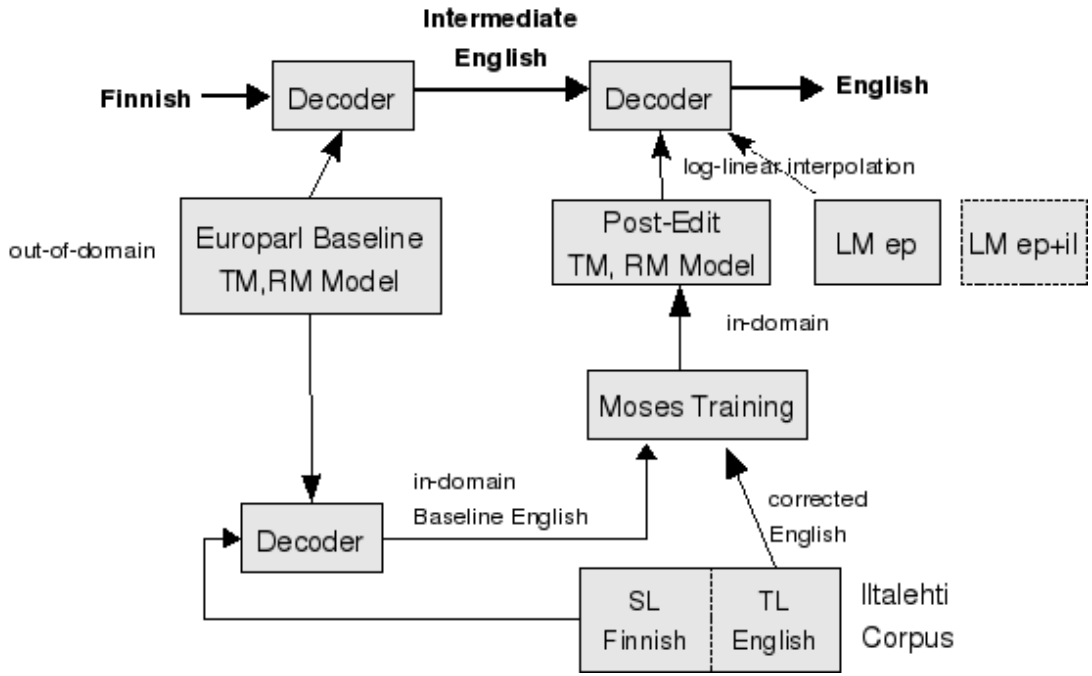
- **I1** combined corpus translation model (TM) and reordering model (RM)
- **I2** I1 + combined corpus LM
- **I3** I1 + linear LM interpolation

Post-edit Method

Two step translation:

1. baseline
2. post-edit model

Corrects the baseline translation to match the reference.



Experiments

- **P1** combined corpus translation model (TM) and reordering model (RM)
- **P2** P1 + combined corpus LM
- **P3** P1 + linear LM interpolation

Evaluation

Automatic machine translation measure

- BLEU: closeness to the human reference is rewarded
- Weighted overlap of n-grams between candidate and reference translation

Translation/language model training

- Very little in-domain training data $\rightarrow$ 10 fold cross-validation
- Bootstrap resampling to improve the BLEU estimate

System ranking

- Pairwise comparison of bootstrapped scores (Zhang et al. 2004)
- Wilcoxon signed rank test

Results

Id	Description	Data			Training	Testing		
		TM	RM	LM		cross-validation		bootstrap
						mean	interval	interval
B	baseline, no adaptaion	ep	ep	ep	16.49	16.43	[15.08, 17.79]	[12.64, 20.38]
L1	log-linear LM combination	ep	ep	ep, il	20.92	13.28	[11.89, 14.68]	[9.91, 16.88]
L2	combined corpus LM	ep	ep	ep+il	20.50	17.25	[15.78, 18.72]	[13.33, 21.40]
L3	linear LM interpolation	ep	ep	ep*il	19.79	14.86	[13.79, 15.93]	[11.42, 18.45]
L4	in-domain LM only	ep	ep	il	20.29	10.77	[9.45, 12.08]	[7.808, 13.87]
C1	combined corpus TM/RM	ep+il	ep+il	ep	48.92	21.41	[19.58, 23.23]	[16.79, 26.32]
C2	+combined corpus LM	ep+il	ep+il	ep+il	55.70	22.41	[20.55, 24.28]	[17.50, 27.57]
C3	+linear LM interpolation	ep+il	ep+il	ep*il	56.19	21.23	[19.73, 22.73]	[16.57, 26.20]
I1	log-linear TM combination	ep, il	ep	ep	62.92	23.75	[21.87, 25.64]	[18.77, 29.04]
I2	+combined corpus LM	ep, il	ep	ep+il	68.98	24.76	[22.49, 27.03]	[19.52, 30.39]
I3	+linear LM interpolation	ep, il	ep	ep*il	69.89	23.43	[21.41, 25.44]	[18.28, 29.08]
P1	post-edit TM/RM	pec	pec	ep	57.75	22.74	[21.24, 24.24]	[17.52, 28.48]
P2	+combined corpus LM	pec	pec	ep+il	61.02	24.05	[22.35, 25.75]	[18.47, 30.01]
P3	+linear LM interpolation	pec	pec	ep*il	61.23	23.49	[21.81, 25.16]	[17.99, 29.35]

Ranking Results

- Bootstrap method

  $(C2, I2, P2) > (L2, B)$

- Wilcoxon signed-rank test

  $P2 > L2 > B$

Id	Description
ep	Europarl (Finnish,English) corpus
il	Iltalehti (Finnish,English) corpus
pec	Post-edit corrections (English,English)
ep+il	One model trained on combined corpora
ep,il	Log-linear combination of models
ep*il	Linear interpolation of models
B	Baseline translation system
Ln	Language model adaptation only
Cn	Adaptation with data combination
In	Adaptation with model interpolation
Pn	Adaptation with post-editing

Summary

- Adaptation methods significantly improve translation performance
- language model adaptation
  (combined corpus LM or linear LM interpolation)
- Log-linear TM combination or post-edit TM/RM methods give best results (translation performance and model training complexity)

log-linear TM combination

- more flexible than post-edit method
- easy combination of several domains
- only models for in-domain data need training
- requires the baseline models

post-edit TM/RM

- easy to add SMT on top of other approaches
- no need for the baseline models
- additional translation step
- translation quality limited by baseline system

Summary

Shortcomings of our experiments

- No parameter tuning (minimum error rate training)
- Very short sentences
- Very small in-domain corpus

$\Rightarrow$ The results might not be generalizable

Thank You