Adaptation of Machine Translation to Specific Domains and Applications

Pavel Pecina

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague
June 7, 2017

Computational Linguistics

- The scientific study of human language from a computational perspective.
- **Subfields**: linguistic analysis of text, document analysis, speech recognition, dialog systems, information retrieval and extraction, machine translation, ...

My Research Area

Computational Linguistics

- The scientific study of human language from a computational perspective.
- **Subfields**: linguistic analysis of text, document analysis, speech recognition, dialog systems, information retrieval and extraction, machine translation, ...

Machine Translation

- Translating text in one human language into another by computer software.
- Application areas:
  - High quality translation
  - Computer-assisted translation
  - Online communication across languages
  - Cross-lingual information retrieval

My Research Area

Computational Linguistics

- The scientific study of human language from a computational perspective.
- **Subfields**: linguistic analysis of text, document analysis, speech recognition, dialog systems, information retrieval and extraction, machine translation, ...

Machine Translation

- Translating text in one human language into another by computer software.
- Application areas:
  - High quality translation
  - Computer-assisted translation
  - Online communication across languages
  - Cross-lingual information retrieval

Machine Translation Approaches

Rule-based:

- Manually created rules and dictionaries map grammatical structures and words from one language to another.

Example-based:

- Input split into phrases which are translated by analogy to preexisting translations in a form of parallel texts.

Statistical:

- Translations generated by statistical models derived from analysis of parallel texts.

Neural:

- Based on deep neural networks trained directly to produce translated texts.

Machine Translation Approaches

Rule-based:

- Manually created rules and dictionaries map grammatical structures and words from one language to another.

Example-based:

- Input split into phrases which are translated by analogy to preexisting translations in a form of parallel texts.

Statistical:

- Translations generated by statistical models derived from analysis of parallel texts.

Neural:

- Based on deep neural networks trained directly to produce translated texts.

Statistical Machine Translation (SMT)

- Translation from F to E generated according to probability distribution $p_{\vec{\theta}}(\mathbf{e}|\mathbf{f})$
  - $\mathbf{f}$: sentence in source language F (e.g. French)
  - $\mathbf{e}$: sentence in target language E (e.g. English)

- $p_{\vec{\theta}}(\mathbf{e}|\mathbf{f})$ approximated by statistical models trained on :
  - a) parallel texts: texts presented in F and E
  - b) monolingual data: texts in E

- Main issues:
  1. Specification of the model
  2. Training model parameters
  3. Finding the best translation

Finding the Best Translation

,

Finding the Best Translation

1. Input sentence: ,

Example: **f**: Morgen gehe ich zur einer Untersuchung ins Krankenhaus .

Finding the Best Translation

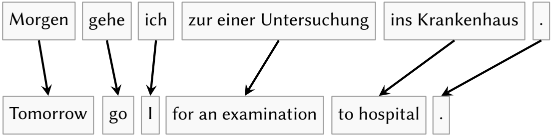
1. Input sentence: segmented into phrases, ,

Example: f: | Morgen | gehe | ich | zur einer Untersuchung | ins Krankenhaus | . |

Finding the Best Translation

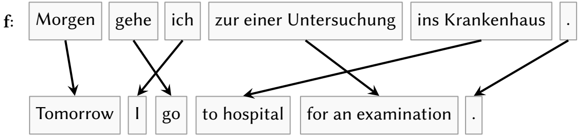
1. Input sentence: segmented into phrases, that are translated,

Example: f:



1. Input sentence: segmented into phrases, that are translated, and reordered

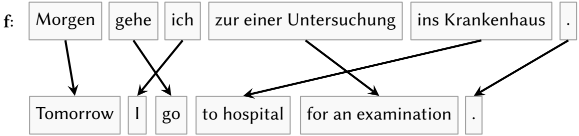
Example:



Finding the Best Translation

1. Input sentence: segmented into phrases, that are translated, and reordered

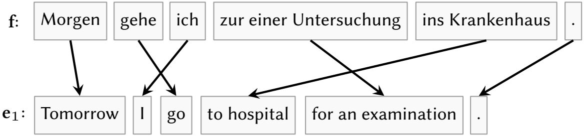
Example:



2. Multiple ways to segment, translate, reorder → multiple hypotheses:

1. Input sentence: segmented into phrases, that are translated, and reordered

Example:

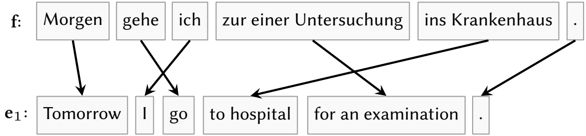


2. Multiple ways to segment, translate, reorder → multiple hypotheses:

Finding the Best Translation

1. Input sentence: segmented into phrases, that are translated, and reordered

Example:



2. Multiple ways to segment, translate, reorder → multiple hypotheses:

e_2: | Tomorrow | I | will go | to hospital | for an examination | . |

Finding the Best Translation

1. Input sentence: segmented into phrases, that are translated, and reordered

Example:

f: | Morgen | gehe | ich | zur einer Untersuchung | ins Krankenhaus | . |

$\mathbf{e}_1$: | Tomorrow | I | go | to hospital | for an examination | . |

2. Multiple ways to segment, translate, reorder → multiple hypotheses:

$\mathbf{e}_2$: | Tomorrow | I | will go | to hospital | for an examination | . |

$\mathbf{e}_3$: | Tomorrow | I | am going | to hospital | for an appointment | . |

Finding the Best Translation

1. Input sentence: segmented into phrases, that are translated, and reordered

Example:

f:	Morgen	gehe	ich	zur einer Untersuchung	ins Krankenhaus	.
$\mathbf{e}_1$:	Tomorrow	I	go	to hospital	for an examination	.

2. Multiple ways to segment, translate, reorder $\rightarrow$ multiple hypotheses:

$\mathbf{e}_2$:	Tomorrow	I	will go	to hospital	for an examination	.
$\mathbf{e}_3$:	Tomorrow	I	am going	to hospital	for an appointment	.
$\mathbf{e}_4$:	In the morning	I'm going	for an examination	to hospital	.	

...

Finding the Best Translation

1. Input sentence: segmented into phrases, that are translated, and reordered

Example:

$\mathbf{f}$:	Morgen	gehe	ich	zur einer Untersuchung	ins Krankenhaus	.	
$\mathbf{e}_1$:	Tomorrow	I	go	to hospital	for an examination	.	-0.17

2. Multiple ways to segment, translate, reorder → multiple hypotheses:

$\mathbf{e}_2$:	Tomorrow	I	will go	to hospital	for an examination	.	-0.15
$\mathbf{e}_3$:	Tomorrow	I	am going	to hospital	for an appointment	.	-0.19
$\mathbf{e}_4$:	In the morning	I'm going	for an examination	to hospital	.		-0.23

...

3. Each hypothesis scored by model: $s(\mathbf{e}, \mathbf{f})$

Finding the Best Translation

1. Input sentence: segmented into phrases, that are translated, and reordered

Example:

f: | Morgen | gehe | ich | zur einer Untersuchung | ins Krankenhaus | . |

$\mathbf{e}_1$: | Tomorrow | I | go | to hospital | for an examination | . | -0.17

2. Multiple ways to segment, translate, reorder $\rightarrow$ multiple hypotheses:

$\mathbf{e}_{\text{best}} = \mathbf{e}_2$:	Tomorrow	I	will go	to hospital	for an examination	.	-0.15
$\mathbf{e}_3$:	Tomorrow	I	am going	to hospital	for an appointment	.	-0.19
$\mathbf{e}_4$:	In the morning	I'm going	for an examination	to hospital	.		-0.23

...

3. Each hypothesis scored by model: $s(\mathbf{e}, \mathbf{f})$

4. The highest-scored hypothesis selected as the best translation: $\mathbf{e}_{\text{best}}$

SMT Model

SMT Model

- Given a foreign sentence $\mathbf{f}$, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ p(\mathbf{e}|\mathbf{f})$$

SMT Model

- Given a foreign sentence $\mathbf{f}$, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ p(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}}$$

SMT Model

- Given a foreign sentence $\mathbf{f}$, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e} \in \text{English}} p(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e} \in \text{English}} \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}} = \arg\max_{\mathbf{e} \in \text{English}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

SMT Model

- Given a foreign sentence **f**, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}\in\text{English}} p(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}\in\text{English}} \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}} = \arg\max_{\mathbf{e}\in\text{English}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

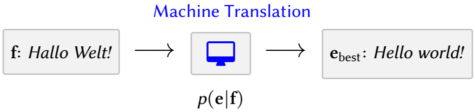
= Noisy channel model motivated by code breaking in information theory:

SMT Model

- Given a foreign sentence $\mathbf{f}$, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ p(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

= Noisy channel model motivated by code breaking in information theory:



SMT Model

- Given a foreign sentence **f**, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ p(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

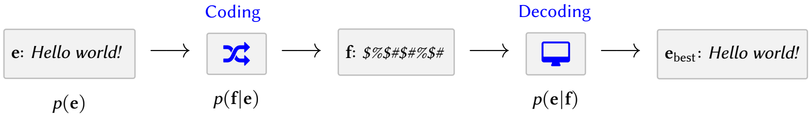
= Noisy channel model motivated by code breaking in information theory:

SMT Model

- Given a foreign sentence $\mathbf{f}$, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e}\in\text{English}}{\arg\max}\; p(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e}\in\text{English}}{\arg\max}\; \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}} = \underset{\mathbf{e}\in\text{English}}{\arg\max}\; p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

= Noisy channel model motivated by code breaking in information theory:

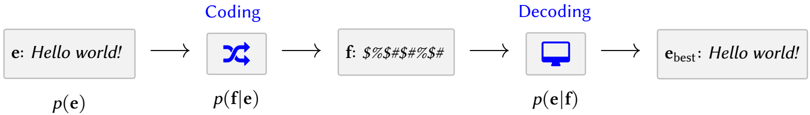


SMT Model

- Given a foreign sentence $\mathbf{f}$, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e}\in\text{English}}{\arg\max}\ p(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e}\in\text{English}}{\arg\max}\ \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}} = \underset{\mathbf{e}\in\text{English}}{\arg\max}\ p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

= Noisy channel model motivated by code breaking in information theory:



- Translation hypothesis $\mathbf{e}$ of the source sentence $\mathbf{f}$ scored by:

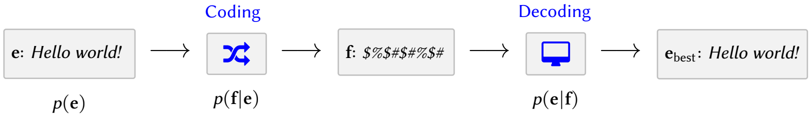
$$s(\mathbf{e}, \mathbf{f}) = p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

SMT Model

- Given a foreign sentence $\mathbf{f}$, find its translation $\mathbf{e}_{\text{best}}$:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e}\in\text{English}}{\arg\max}\; p(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e}\in\text{English}}{\arg\max}\; \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{\cancel{p(\mathbf{f})}} = \underset{\mathbf{e}\in\text{English}}{\arg\max}\; p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

= Noisy channel model motivated by code breaking in information theory:



- Translation hypothesis $\mathbf{e}$ of the source sentence $\mathbf{f}$ scored by:

$$s(\mathbf{e},\mathbf{f}) = p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

- Basic model components:
  - Translation model $p(\mathbf{f}|\mathbf{e})$ – modeling adequacy of translation $\mathbf{e} \rightarrow \mathbf{f}$
  - Language model $p(\mathbf{e})$ – modeling fluency of $\mathbf{e}$

Training Model Parameters

Translation model: $p(\mathbf{f}|\mathbf{e})$

- Approximates probability of $\mathbf{e} \rightarrow \mathbf{f}$ from parallel texts

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"rýmu"}) = \frac{\#(\mathbf{e}=\text{"rýmu"}, \mathbf{f}=\text{"a cold"})}{\#(\mathbf{e}=\text{"rýmu"})} = 0.020389$$

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"nachlazení"}) = \frac{\#(\mathbf{f}=\text{"a cold"}, \mathbf{e}=\text{"nachlazení"})}{\#(\mathbf{e}=\text{"nachlazení"})} = 0.023231$$

Training Model Parameters

Translation model: $p(\mathbf{f}|\mathbf{e})$

- Approximates probability of $\mathbf{e} \rightarrow \mathbf{f}$ from parallel texts

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"rýmu"}) = \frac{\#(\mathbf{e}=\text{"rýmu"}, \mathbf{f}=\text{"a cold"})}{\#(\mathbf{e}=\text{"rýmu"})} = 0.020389$$

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"nachlazení"}) = \frac{\#(\mathbf{f}=\text{"a cold"}, \mathbf{e}=\text{"nachlazení"})}{\#(\mathbf{e}=\text{"nachlazení"})} = 0.023231$$

Language model: $p(\mathbf{e})$

- Approximates probability of $\mathbf{e}$ from monolingual texts

$$p(\mathbf{e}=\text{"dostal nachlazení"}) = \frac{\#(\mathbf{e}=\text{"dostal nachlazení"})}{N} = 0.072 \cdot 10^{-6}$$

$$p(\mathbf{e}=\text{"dostal rýmu"}) = \frac{\#(\mathbf{e}=\text{"dostal rýmu"})}{N} = 0.351 \cdot 10^{-6}$$

Training Model Parameters

Translation model: $p(\mathbf{f}|\mathbf{e})$

- Approximates probability of $\mathbf{e} \to \mathbf{f}$ from parallel texts

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"rýmu"}) = \frac{\#(\mathbf{e}=\text{"rýmu"}, \mathbf{f}=\text{"a cold"})}{\#(\mathbf{e}=\text{"rýmu"})} = 0.020389$$

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"nachlazení"}) = \frac{\#(\mathbf{f}=\text{"a cold"}, \mathbf{e}=\text{"nachlazení"})}{\#(\mathbf{e}=\text{"nachlazení"})} = 0.023231$$

Language model: $p(\mathbf{e})$

- Approximates probability of $\mathbf{e}$ from monolingual texts

$$p(\mathbf{e}=\text{"dostal nachlazení"}) = \frac{\#(\mathbf{e}=\text{"dostal nachlazení"})}{N} = 0.072 \cdot 10^{-6}$$

$$p(\mathbf{e}=\text{"dostal rýmu"}) = \frac{\#(\mathbf{e}=\text{"dostal rýmu"})}{N} = 0.351 \cdot 10^{-6}$$

Example:

$$\begin{array}{llll} \mathbf{f}: & \boxed{\text{Peter}} & \boxed{\text{got}} & \boxed{\text{a cold}} \\ & \downarrow & \downarrow & \downarrow \\ \mathbf{e}_1: & \boxed{\text{Petr}} & \boxed{\text{dostal}} & \boxed{\text{nachlazení}} \end{array} \qquad \begin{array}{llll} \mathbf{f}: & \boxed{\text{Peter}} & \boxed{\text{got}} & \boxed{\text{a cold}} \\ & \downarrow & \downarrow & \downarrow \\ \mathbf{e}_2: & \boxed{\text{Petr}} & \boxed{\text{dostal}} & \boxed{\text{rýmu}} \end{array}$$

Translation model: $p(\mathbf{f}|\mathbf{e})$

- Approximates probability of $\mathbf{e} \rightarrow \mathbf{f}$ from parallel texts

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"rýmu"}) = \frac{\#(\mathbf{e}=\text{"rýmu"}, \mathbf{f}=\text{"a cold"})}{\#(\mathbf{e}=\text{"rýmu"})} = 0.020389$$

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"nachlazení"}) = \frac{\#(\mathbf{f}=\text{"a cold"}, \mathbf{e}=\text{"nachlazení"})}{\#(\mathbf{e}=\text{"nachlazení"})} = 0.023231$$

Language model: $p(\mathbf{e})$

- Approximates probability of $\mathbf{e}$ from monolingual texts

$$p(\mathbf{e}=\text{"dostal nachlazení"}) = \frac{\#(\mathbf{e}=\text{"dostal nachlazení"})}{N} = 0.072 \cdot 10^{-6}$$

$$p(\mathbf{e}=\text{"dostal rýmu"}) = \frac{\#(\mathbf{e}=\text{"dostal rýmu"})}{N} = 0.351 \cdot 10^{-6}$$

Example:

$\mathbf{f}$:	Peter	got	a cold
$\mathbf{e}_1$:	Petr	dostal	nachlazení

$$s(\mathbf{e}_1, \mathbf{f}) < s(\mathbf{e}_2, \mathbf{f})$$

$\mathbf{f}$:	Peter	got	a cold
$\mathbf{e}_2$:	Petr	dostal	rýmu

Translation model: $p(\mathbf{f}|\mathbf{e})$

- Approximates probability of $\mathbf{e} \rightarrow \mathbf{f}$ from parallel texts ($\sim 10^7$ sentence pairs)

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"rýmu"}) = \frac{\#(\mathbf{e}=\text{"rýmu"}, \mathbf{f}=\text{"a cold"})}{\#(\mathbf{e}=\text{"rýmu"})} = 0.020389$$

$$p(\mathbf{f}=\text{"a cold"}|\mathbf{e}=\text{"nachlazení"}) = \frac{\#(\mathbf{f}=\text{"a cold"}, \mathbf{e}=\text{"nachlazení"})}{\#(\mathbf{e}=\text{"nachlazení"})} = 0.023231$$

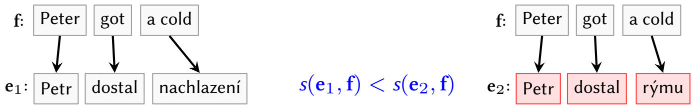
Language model: $p(\mathbf{e})$

- Approximates probability of $\mathbf{e}$ from monolingual texts ($\sim 10^9$ words)

$$p(\mathbf{e}=\text{"dostal nachlazení"}) = \frac{\#(\mathbf{e}=\text{"dostal nachlazení"})}{N} = 0.072 \cdot 10^{-6}$$

$$p(\mathbf{e}=\text{"dostal rýmu"}) = \frac{\#(\mathbf{e}=\text{"dostal rýmu"})}{N} = 0.351 \cdot 10^{-6}$$

Example:



SMT Model cont.

- Noisy channel model

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\; p(\mathbf{f}|\mathbf{e})\; p(\mathbf{e}) \qquad \cdots$$

SMT Model cont.

- Noisy channel model extended to allow:
  1. additional components: reordering model $r(\mathbf{e}, \mathbf{f})$, length model $l(\mathbf{e}, \mathbf{f})$, ...

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\; p(\mathbf{f}|\mathbf{e})\; p(\mathbf{e})\; r(\mathbf{e}, \mathbf{f})\; l(\mathbf{e}, \mathbf{f}) \quad \cdots$$

SMT Model cont.

- Noisy channel model extended to allow:
  1. additional components: reordering model $r(\mathbf{e}, \mathbf{f})$, length model $l(\mathbf{e}, \mathbf{f})$, ...
  2. component weights: λ_i

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\ p(\mathbf{f}|\mathbf{e})^{\lambda_1} p(\mathbf{e})^{\lambda_2} r(\mathbf{e}, \mathbf{f})^{\lambda_3} l(\mathbf{e}, \mathbf{f})^{\lambda_4} \cdots$$

SMT Model cont.

- Noisy channel model extended to allow:
  1. additional components: reordering model $r(\mathbf{e}, \mathbf{f})$, length model $l(\mathbf{e}, \mathbf{f})$, ...
  2. component weights: λ_i

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max}\; p(\mathbf{f}|\mathbf{e})^{\lambda_1} p(\mathbf{e})^{\lambda_2} r(\mathbf{e}, \mathbf{f})^{\lambda_3} l(\mathbf{e}, \mathbf{f})^{\lambda_4} \cdots$$

- Final model:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e} \in \text{English}}{\arg\max} \prod_{i=1}^{n} h_i(e, f)^{\lambda_i} = \underset{\mathbf{e} \in \text{English}}{\arg\max} \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e}, \mathbf{f})$$

SMT Model cont.

- Noisy channel model extended to allow:
  1. additional components: reordering model $r(\mathbf{e},\mathbf{f})$, length model $l(\mathbf{e},\mathbf{f})$, ...
  2. component weights: λ_i

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e}\in\text{English}}{\arg\max}\; p(\mathbf{f}|\mathbf{e})^{\lambda_1} p(\mathbf{e})^{\lambda_2} r(\mathbf{e},\mathbf{f})^{\lambda_3} l(\mathbf{e},\mathbf{f})^{\lambda_4} \cdots$$

- Final model:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e}\in\text{English}}{\arg\max} \prod_{i=1}^{n} h_i(e,f)^{\lambda_i} = \underset{\mathbf{e}\in\text{English}}{\arg\max} \boxed{\sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e},\mathbf{f})} \leftarrow$$

log-linear model
$s(\mathbf{e},\mathbf{f})$

SMT Model cont.

- Noisy channel model extended to allow:
  1. additional components: reordering model $r(\mathbf{e},\mathbf{f})$, length model $l(\mathbf{e},\mathbf{f})$, ...
  2. component weights: λ_i

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e}\in\text{English}}{\text{arg max}}\; p(\mathbf{f}|\mathbf{e})^{\lambda_1} p(\mathbf{e})^{\lambda_2} r(\mathbf{e},\mathbf{f})^{\lambda_3} l(\mathbf{e},\mathbf{f})^{\lambda_4} \cdots$$

- Final model:

$$\mathbf{e}_{\text{best}} = \underset{\mathbf{e}\in\text{English}}{\text{arg max}} \prod_{i=1}^{n} h_i(e,f)^{\lambda_i} = \underset{\mathbf{e}\in\text{English}}{\text{arg max}} \boxed{\sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e},\mathbf{f})} \leftarrow \underset{s(\mathbf{e},\mathbf{f})}{\text{log-linear model}}$$

Weights λ_i – model hyperparameters

- Tuned by Minimum Error Rate Training (MERT) to maximize translation quality on development set of parallel texts ($\sim 10^3$ sentence pairs).
- Traslation quality measured as similarity to human translation (e.g. BLEU).

My Work:
Adaptation of Statistical Machine Translation

The Problem

Theory:

- SMT training $\times$ test data $\sim$ the same distribution

The Problem

Theory:

- SMT training $\times$ test data $\sim$ the same distribution

Practice:

- SMT trained on: news, parliamentary proceedings, novels, movie subtitles $\rightarrow$ general (non-specific) domain.

$\Rightarrow$ Translation quality decreases when translating text from specific domains:

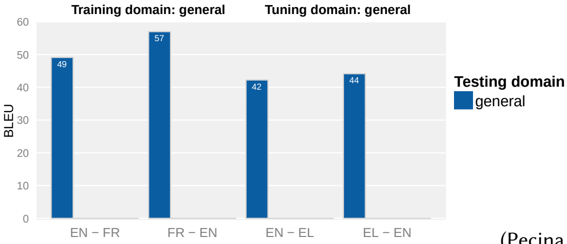
Theory:

- SMT training $\times$ test data $\sim$ the same distribution

Practice:

- SMT trained on: news, parliamentary proceedings, novels, movie subtitles $\rightarrow$ general (non-specific) domain.

$\Rightarrow$ Translation quality decreases when translating text from specific domains:



(Pecina et al., 2012)

The Problem

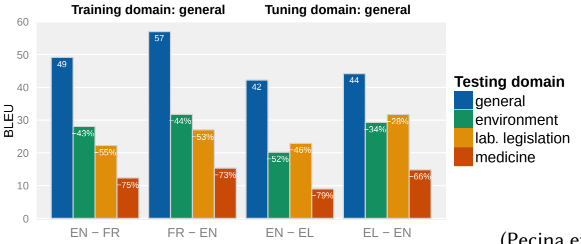
Theory:

- SMT training $\times$ test data $\sim$ the same distribution

Practice:

- SMT trained on: news, parliamentary proceedings, novels, movie subtitles $\rightarrow$ general (non-specific) domain.

$\Rightarrow$ Translation quality decreases when translating text from specific domains:



(Pecina et al., 2012)

The Problem

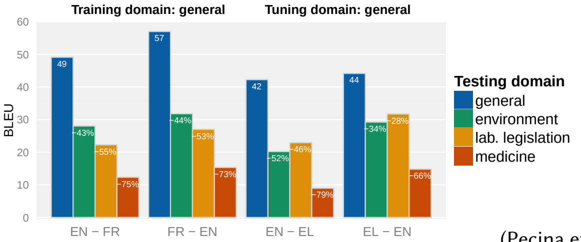
Theory:

- SMT training $\times$ test data $\sim$ the same distribution

Practice:

- SMT trained on: news, parliamentary proceedings, novels, movie subtitles $\rightarrow$ general (non-specific) domain.

$\Rightarrow$ Translation quality decreases when translating text from specific domains:



(Pecina et al., 2012)

Problem:

- Improving translation for specific domains

P. Pecina, A. Toral, V. Papavassiliou, P. Prokopidis, J. van Genabith: Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A Case Study, *Proceedings of the 16th Annual Conference of EAMT*, 145-152. Trento, Italy, 2012..

Research Goals

1. Adaptation of SMT to specific-domains
2. Adaptation of SMT to cross-lingual information retrieval
3. Acquisition of domain-specific data for SMT

Research Goals

1. Adaptation of SMT to specific-domains
2. Adaptation of SMT to cross-lingual information retrieval
3. Acquisition of domain-specific data for SMT

Research Goals

1. Adaptation of SMT to specific-domains
2. Adaptation of SMT to cross-lingual information retrieval
3. Acquisition of domain-specific data for SMT

The work conducted within three research projects funded by EU:

PANACEA — FP7, 2010–2012

khresmoi (MEDICAL INFORMATION ANALYSIS & RETRIEVAL) — FP7, 2012–2014

KCONNECT — H2020, 2015–2017

1. Adaptation of SMT to Specific Domains

Domain Adaptation of SMT

Task:

- Adaptation of SMT systems trained on general domain data to a specific domain for which no or only limited data is available.

What can be improved?

1. Hyperparameters λ_i tuning (Pecina et al., 2012)
2. Language $p(\mathbf{e})$ and translation model $p(\mathbf{f}|\mathbf{e})$ training (Pecina et al., 2015)
3. Training data preprocessing (Toral et al., 2015)

P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

P. Pecina, A. Toral, V. Papavassiliou, P. Prokopidis, A. Tamchyna, A. Way, J. van Genabith: Domain Adaptation of Statistical MT with Domain-focused Web Crawling, *Language Resources and Evaluation*, 49(1), 147–193. Springer, 2015. (**IF 0.975**).

A. Toral, P. Pecina, L. Wang, J. van Genabith: Linguistically-augmented Perplexity-based Data Selection for Language Models, *Computer Speech and Language, Special Issue on Hybrid Machine Translation: Integration of Linguistics and Statistics*, 32(1), pp. 11–26. Elsevier, 2015. (**IF 1.324**).

Domain Adaptation of SMT

Task:

- Adaptation of SMT systems trained on general domain data to a specific domain for which no or only limited data is available.

What can be improved?

1. Hyperparameters λ_i tuning (Pecina et al., 2012)
2. Language $p(\mathbf{e})$ and translation model $p(\mathbf{f}|\mathbf{e})$ training (Pecina et al., 2015)
3. Training data preprocessing (Toral et al., 2015)

- P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).
- P. Pecina, A. Toral, V. Papavassiliou, P. Prokopidis, A. Tamchyna, A. Way, J. van Genabith: Domain Adaptation of Statistical MT with Domain-focused Web Crawling, *Language Resources and Evaluation*, 49(1), 147–193. Springer, 2015. (**IF 0.975**).
- A. Toral, P. Pecina, L. Wang, J. van Genabith: Linguistically-augmented Perplexity-based Data Selection for Language Models, *Computer Speech and Language, Special Issue on Hybrid Machine Translation: Integration of Linguistics and Statistics*, 32(1), pp. 11–26. Elsevier, 2015. (**IF 1.324**).

Hyperparameter Tuning

Model:

- $s(\mathbf{e}, \mathbf{f}) = \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e}, \mathbf{f})$, λ_i tuned by MERT on development data.

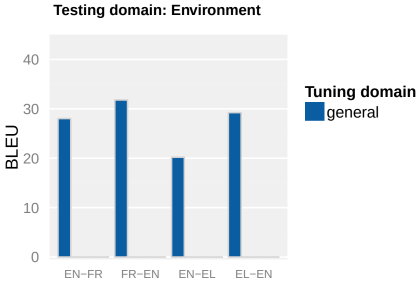
Hyperparameter Tuning

Model:

- $s(\mathbf{e}, \mathbf{f}) = \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e}, \mathbf{f})$, λ_i tuned by MERT on development data.

Standard approach:

- tuning on general domain
- h_i and λ_i optimized on general domain



Hyperparameter Tuning

Model:

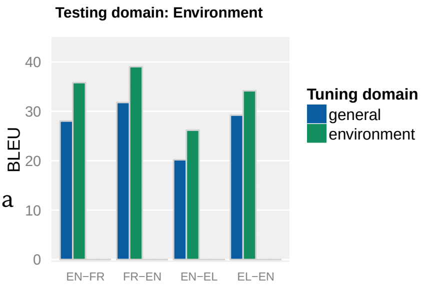
- $s(\mathbf{e},\mathbf{f}) = \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e},\mathbf{f})$, λ_i tuned by MERT on development data.

Standard approach:

- tuning on general domain
- h_i and λ_i optimized on general domain

Ideal approach:

- tuning on test domain development data
- BLEU: +33%



Hyperparameter Tuning

Model:

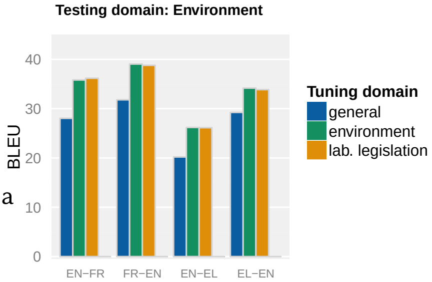
- $s(\mathbf{e},\mathbf{f}) = \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e},\mathbf{f})$, λ_i tuned by MERT on development data.

Standard approach:

- tuning on general domain
- h_i and λ_i optimized on general domain

Ideal approach:

- tuning on test domain development data
- BLEU: +33%



Our approach (Pecina et al., 2012):

- cross-domain tuning
- when no test-domain development data is available
- BLEU: +30%

P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

Hyperparameter Tuning

Model:

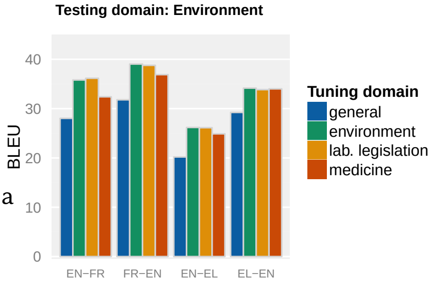
- $s(\mathbf{e},\mathbf{f}) = \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e},\mathbf{f})$, λ_i tuned by MERT on development data.

Standard approach:

- tuning on general domain
- h_i and λ_i optimized on general domain

Ideal approach:

- tuning on test domain development data
- BLEU: +33%



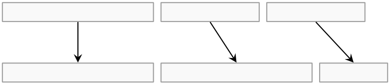
Our approach (Pecina et al., 2012):

- cross-domain tuning
- when no test-domain development data is available
- BLEU: +30%

P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

Effect of Tuning for Specific Domains

1. General-domain-tuned translation:

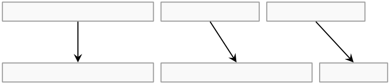


- longer phrases
- infrequently reordered

P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

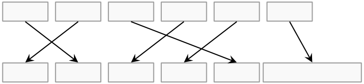
Effect of Tuning for Specific Domains

1. General-domain-tuned translation:



   - longer phrases
   - infrequently reordered

2. Specific-domain-tuned translation:

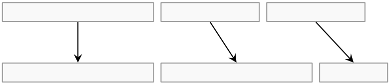


   - shorter phrases
   - heavily reordered

P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

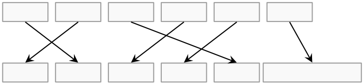
Effect of Tuning for Specific Domains

1. General-domain-tuned translation:



- longer phrases
- infrequently reordered

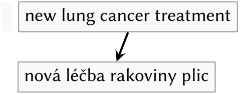
2. Specific-domain-tuned translation:



- shorter phrases
- heavily reordered

Example:

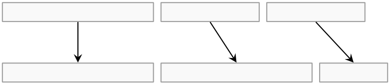
- Ideal translation:



P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

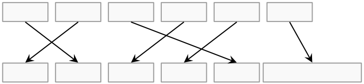
Effect of Tuning for Specific Domains

1. General-domain-tuned translation:



- longer phrases
- infrequently reordered

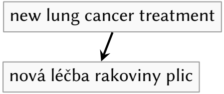
2. Specific-domain-tuned translation:



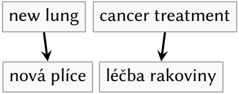
- shorter phrases
- heavily reordered

Example:

- Ideal translation:



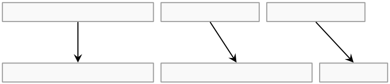
- General-domain-tuned translation:



P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

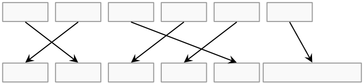
Effect of Tuning for Specific Domains

1. General-domain-tuned translation:



- longer phrases
- infrequently reordered

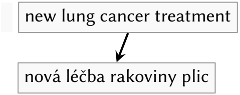
2. Specific-domain-tuned translation:



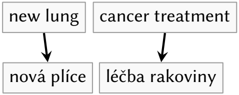
- shorter phrases
- heavily reordered

Example:

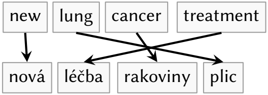
- Ideal translation:



- General-domain-tuned translation:



- Specific-domain-tuned translation:



P. Pecina, A. Toral, J. van Genabith: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation, *Proceedings of the 24th International Conference on Computational Linguistics*, 2209–2224. Mumbai, India, 2012. (**A**).

2. SMT for Cross-lingual Information Retrieval

Cross-Lingual Information Retrieval

Cross-Lingual Information Retrieval

Information Retrieval (IR)

- Searching for relevant documents within a large collection (e.g. web search).

Information Retrieval (IR)

- Searching for relevant documents within a large collection (e.g. web search).

query

bilateral infiltrates and radiography

$\xrightarrow{\textit{retrieval}}$

top documents

1: *0.92, doc_8343672,*
2: *0.85, doc_4329433,*
3: *0.78, doc_1312294,*
⋮
9: *0.42, doc_0343297,*
10: *0.33, doc_9837244,*
⋮

Cross-Lingual Information Retrieval

Information Retrieval (IR)

- Searching for relevant documents within a large collection (e.g. web search).

query in English

bilateral infiltrates and radiography

$\xrightarrow{\textit{retrieval}}$

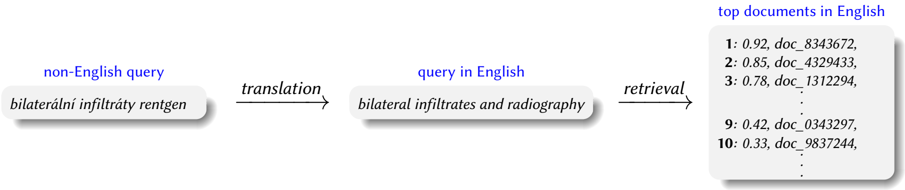
top documents in English

1: *0.92, doc_8343672,*
2: *0.85, doc_4329433,*
3: *0.78, doc_1312294,*
⋮
9: *0.42, doc_0343297,*
10: *0.33, doc_9837244,*
⋮

Cross-Lingual Information Retrieval

Information Retrieval (IR)

- Searching for relevant documents within a large collection (e.g. web search).



Cross-Lingual Information Retrieval (CLIR)

- Query language differs from the document language.
- Useful for:
  a) search in multilingual collections
  b) foreign speakers

Cross-Lingual Information Retrieval

Information Retrieval (IR)

- Searching for relevant documents within a large collection (e.g. web search).

Cross-Lingual Information Retrieval (CLIR)

- Query language differs from the document language.
- Useful for:
  - a) search in multilingual collections
  - b) foreign speakers

Evaluation:

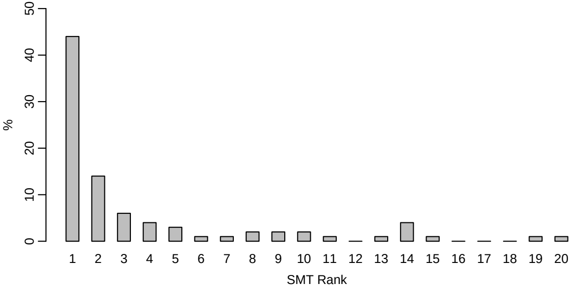
- P@10 (ratio of relevant documents among top 10)

SMT for Query Translation

- **Standard approach:** use the single best query translation by SMT
- **Problem:**
  - SMT trained towards translation quality (e.g. BLEU).
  - CLIR evaluated based on retrieval quality (e.g. P@10).
  - Translation quality does not correlate well with retrieval quality.

SMT for Query Translation

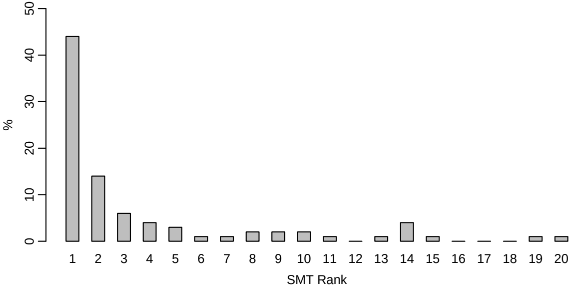
- **Standard approach:** use the single best query translation by SMT
- **Problem:**
  - SMT trained towards translation quality (e.g. BLEU).
  - CLIR evaluated based on retrieval quality (e.g. P@10).
  - Translation quality does not correlate well with retrieval quality.



Distribution of the most IR-useful translations among top 20 SMT translations

SMT for Query Translation

- **Standard approach:** use the single best query translation by SMT
- **Problem:**
  - SMT trained towards translation quality (e.g. BLEU).
  - CLIR evaluated based on retrieval quality (e.g. P@10).
  - Translation quality does not correlate well with retrieval quality.



Distribution of the most IR-useful translations among top 20 SMT translations

- **Our approach**: reranking multiple SMT translations (Saleh & Pecina, 2016)

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Translation Reranking

Query Translation Reranking

query in Czech

bilaterální infiltráty rentgen

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Translation Reranking

query in Czech

bilaterální infiltráty rentgen $\xrightarrow{SMT}$

English translation options	s(e,f)
1: *bilateral infiltrates radiography*	**-0,15**
2: *bilateral infiltrates roentgen*	**-0,19**
⋮	
6: *bilateral infiltrates x-ray*	**-0,21**
⋮	

1. SMT produces multiple translation options (e.g. 20)

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Translation Reranking

query in Czech

bilaterální infiltráty rentgen $\xrightarrow{SMT}$

English translation options	s(e,f)
1: *bilateral infiltrates radiography*	**-0,15**
2: *bilateral infiltrates roentgen*	**-0,19**
⋮	
6: *bilateral infiltrates x-ray*	**-0,21**
⋮	

1. SMT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features:

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Translation Reranking

query in Czech

bilaterální infiltráty rentgen $\xrightarrow{SMT}$

English translation options	P@10
1: *bilateral infiltrates radiography*	**0,62**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
6: *bilateral infiltrates x-ray*	**0,91**
⋮	

1. SMT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features:
3. A regression model trained to predict P@10 for each translation

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Translation Reranking

query in Czech

bilaterální infiltráty rentgen

$\xrightarrow{SMT}$

English translation options — P@10

1: *bilateral infiltrates radiography*	**0,62**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
6: *bilateral infiltrates x-ray*	**0,91**
⋮	

$\xrightarrow{reranking}$

selected English translation

6: *bilateral infiltrates x-ray*	**0,91**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
1: *bilateral infiltrates radiography*	**0,62**
⋮	

1. SMT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features:
3. A regression model trained to predict P@10 for each translation
4. Reranking according to the predicted P@10 scores

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Translation Reranking

query in Czech

bilaterální infiltráty rentgen

$\xrightarrow{SMT}$

English translation options — P@10

1:	*bilateral infiltrates radiography*	**0,62**
2:	*bilateral infiltrates roentgen*	**0,89**
⋮		
6:	*bilateral infiltrates x-ray*	**0,91**
⋮		

$\xrightarrow{reranking}$

selected English translation

6:	*bilateral infiltrates x-ray*	**0,91**
2:	*bilateral infiltrates roentgen*	**0,89**
⋮		
1:	*bilateral infiltrates radiography*	**0,62**
⋮		

1. SMT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features:
3. A regression model trained to predict P@10 for each translation
4. Reranking according to the predicted P@10 scores
5. The highest-scored translation selected

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Translation Reranking

query in Czech

bilaterální infiltráty rentgen

$\xrightarrow{\text{SMT}}$

English translation options — **P@10**

1: *bilateral infiltrates radiography*	**0,62**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
6: *bilateral infiltrates x-ray*	**0,91**
⋮	

$\xrightarrow{\text{reranking}}$

selected English translation

6: *bilateral infiltrates x-ray*	**0,91**
2: *bilateral infiltrates roentgen*	**0,89**
⋮	
1: *bilateral infiltrates radiography*	**0,62**
⋮	

1. SMT produces multiple translation options (e.g. 20)
2. Each translation option represented by a vector of features:
3. A regression model trained to predict P@10 for each translation
4. Reranking according to the predicted P@10 scores
5. The highest-scored translation selected

Features:

- internal SMT features (h_i scores)
- external features: word frequencies from document collection, Wikipedia, …

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Query Reranking Experiments

Evaluated using the CLEF eHealth collection:

- 1 million web-crawled medical documents in English
- 166 medical queries in English, Czech, French, German
- Relevance assesment by medical experts

Query Reranking Experiments

Evaluated using the CLEF eHealth collection:

- 1 million web-crawled medical documents in English
- 166 medical queries in English, Czech, French, German
- Relevance assesment by medical experts

Results (P@10): (Saleh & Pecina, 2016)

	query language		
system	Czech	French	German

Query Reranking Experiments

Evaluated using the CLEF eHealth collection:

- 1 million web-crawled medical documents in English
- 166 medical queries in English, Czech, French, German
- Relevance assesment by medical experts

Results (P@10): (Saleh & Pecina, 2016)

	query language		
system	Czech	French	German
Baseline (SMT)	45.61	47.73	42.42

Query Reranking Experiments

Evaluated using the CLEF eHealth collection:

- 1 million web-crawled medical documents in English
- 166 medical queries in English, Czech, French, German
- Relevance assesment by medical experts

Results (P@10): (Saleh & Pecina, 2016)

	query language		
system	Czech	French	German
Baseline (SMT)	45.61	47.73	42.42
Reranking	50.15	51.06	45.30

Query Reranking Experiments

Evaluated using the CLEF eHealth collection:

- 1 million web-crawled medical documents in English
- 166 medical queries in English, Czech, French, German
- Relevance assesment by medical experts

Results (P@10): (Saleh & Pecina, 2016)

	query language		
system	Czech	French	German
Baseline (SMT)	45.61	47.73	42.42
Reranking	50.15	51.06	45.30
Monolingual (English)	50.30	50.30	50.30

Query Reranking Experiments

Evaluated using the CLEF eHealth collection:

- 1 million web-crawled medical documents in English
- 166 medical queries in English, Czech, French, German
- Relevance assesment by medical experts

Results (P@10): (Saleh & Pecina, 2016)

	query language		
system	Czech	French	German
Baseline (SMT)	45.61	47.73	42.42
Reranking	50.15	51.06	45.30
Monolingual (English)	50.30	50.30	50.30
Google Translate	50.91	49.70	49.39
Bing Translator	47.88	48.64	46.52

Query Reranking Experiments

Evaluated using the CLEF eHealth collection:

- 1 million web-crawled medical documents in English
- 166 medical queries in English, Czech, French, German
- Relevance assesment by medical experts

Results (P@10): (Saleh & Pecina, 2016)

	query language		
system	Czech	French	German
Baseline (SMT)	45.61	47.73	42.42
Reranking	50.15	51.06	45.30
Monolingual (English)	50.30	50.30	50.30
Google Translate	50.91	49.70	49.39
Bing Translator	47.88	48.64	46.52

S. Saleh, P. Pecina: Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval, *Proceedings of the 7th International Conference of the CLEF Association*, 54–66. Évora, Portugal, 2016. SCOPUS.

Conclusions

Conclusions and Future Work

Main findings

1. SMT can be effectively adapted to specific domains
2. SMT reranking improves cross-lingual information retrieval.
3. Web is a great source of domain-specific language data.

Conclusions and Future Work

Main findings

1. SMT can be effectively adapted to specific domains
2. SMT reranking improves cross-lingual information retrieval.
3. Web is a great source of domain-specific language data.

Future Work

1. Adaptation of Neural Machine Translation (NMT)
    - The problem persists: NMT requires adaptation to specific domains
    - Not much work done in this area so far.
2. Cross-lingual Information Retrieval
    - Exploitation of cross-lingual word embeddings