LISBOA

UNIVERSIDADE
DE LISBOA

qtleap

Machine Translation for Multilingual Troubleshooting in the IT Domain: A Comparison of Different Strategies

Sanja Štajner, João Rodrigues, Luís Gomes and António Branco

Department of Informatics, Faculty of Sciences

University of Lisbon, Portugal

DMTW, Prague, Czech Republic

September 4, 2015

Outline

- Problem
- Strategies
- Methodology
- Results
- Conclusions

Problem

- English-Portuguese MT is rarely addressed
- No studies addressing this problem for specific domains
- Domain-specific parallel corpora (EN-PT) are scarce

Strategies

1. Adding out-of-domain corpora
2. Adding in-domain bilingual terminology
3. Adding combination of both (out-of-domain corpora and in-domain bilingual terminology)

Focus

- English to Portuguese MT
- Short sentences (user questions followed by answers from an IT technician)
- Continious chats

Corpora

1. **EP** – English to Portuguese Europarl (1,960,407 sentence pairs) as the large out-of-domain corpus

2. **IT1** – An in-domain IT corpus with 2,000 sentence pairs (1,000 questions and 1,000 answers) compiled under the QTLeap project (used for training)

3. **IT2** – An in-domain IT corpus with 1,000 sentence pairs (answers only) compiled under the QTLeap project (used for testing)

4. **TERM** – A parallel corpus of IT terminology (unigrams or multiword expressions), which consists of the Microsoft Terminology Collection (13,030 terms) and a small portion of LibreOffice terminology (995 terms).

Examples

Corpora	Source (EN)	Target (PT)
TERM	arrow key gatekeeper Planning System Database	tecla de seta controlador de chamadas Base de Dados do Sistema de Planeamento
IT1	If your disc is not recognized, try changing the USB port. Which antivirus should I keep, MSE or AVG?	Se o disco não está a ser reconhecido, tente trocar de entrada USB. Qual antivrus devo manter, MSE ou AVG?
IT2	In the Insert menu, select Picture. In the taskbar there is an icon shaped like binoculars, click and type in what you want to search.	No menu inserir selecione Imagem. Na barra de Tarefas há um ícone em forma de binóculos, clique e escreva o que pretende procurar.
EP	Please rise, then, for this minute's silence. You have requested a debate on this subject in the course of the next few days, during this part-session.	Convido-os a levantarem-se para um minuto de silêncio. Os senhores manifestaram o desejo de se proceder a um debate sobre o assunto nos próximos dias, durante este período de sessões.

Experiments

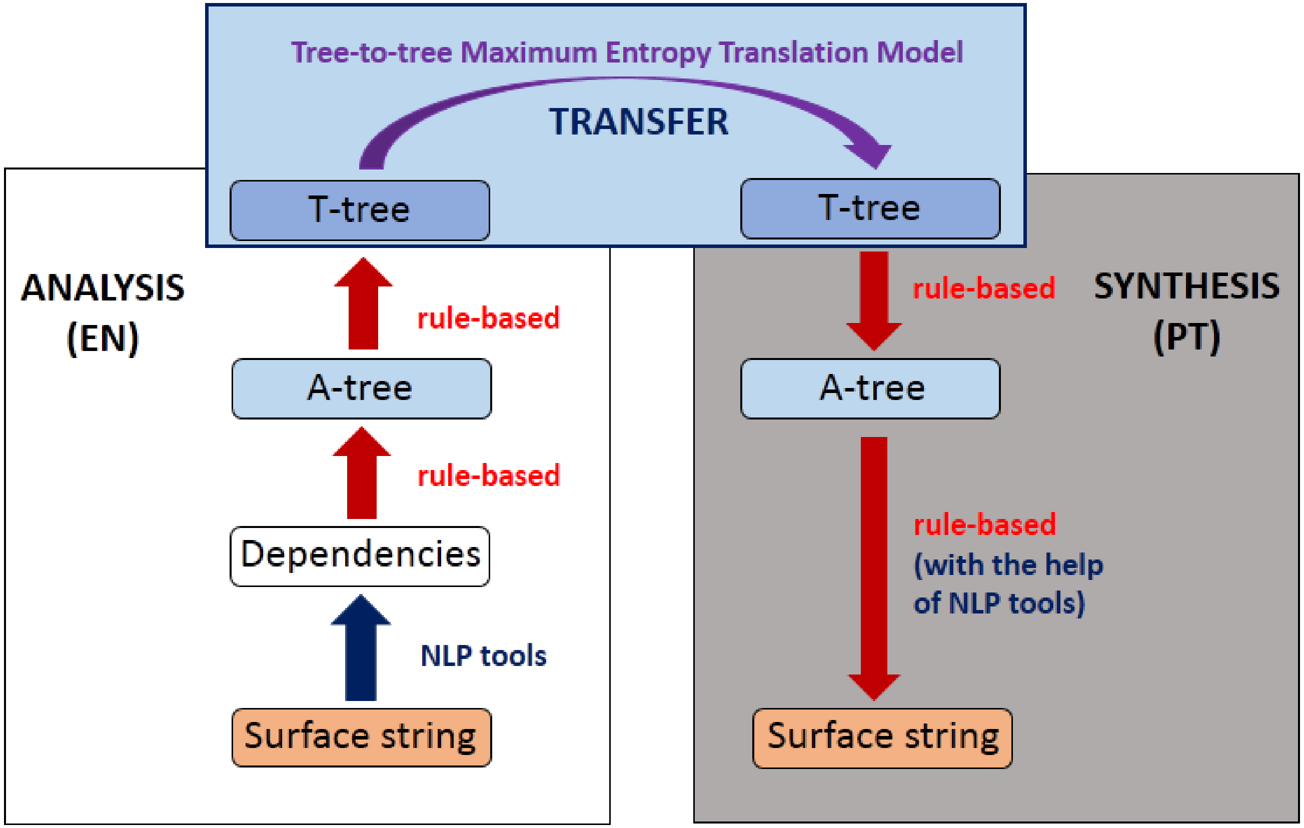
MT Systems :

- A hybrid MT system (TectoMT)
- A standard PBSMT system (Moses)

Training datasets:

- IT+TERM (adding terminology)
- IT+EP1 (adding out-of-domain data)
- IT+EP10
- IT+EP10+TERM (adding both)

TectoMT



Tree-to-tree Maximum Entropy Translation Model

TRANSFER

T-tree

T-tree

ANALYSIS (EN)

rule-based

A-tree

rule-based

Dependencies

NLP tools

Surface string

rule-based

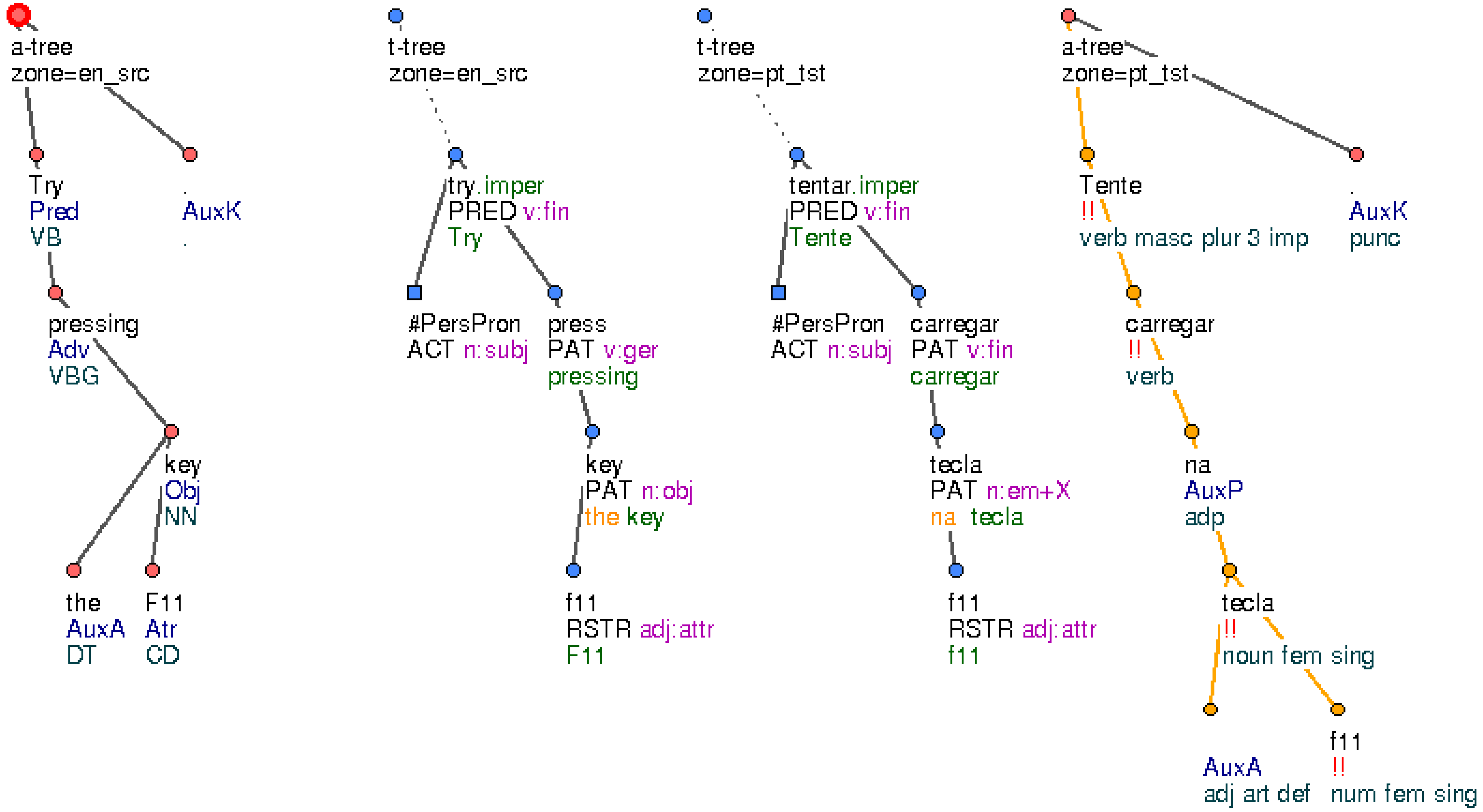
SYNTHESIS (PT)

A-tree

rule-based (with the help of NLP tools)

Surface string

A-tree vs. T-tree



"Try pressing the F11 key." translated into "Tente carregar na tecla f11."

Human Evaluation Parameters

- Scores:
  - Fluency (1 – 4)
  - Adequacy (1 – 4)

1 – very bad
2 – bad
3 – good
4 – very good

- Error Analysis:
  - Orthographic (0 – 2)
  - Morphologic (0 – 2)
  - Syntactic (0 – 2)
  - Semantic (0 – 2)

0 – no errors
1 – one error
2 – two or more errors

Results (Automatic Evaluation)

Experiment	Training			Dev.	Test	Results (BLEU)	
	EP	TERM	IT1	IT1	IT2	TectoMT	PBSMT
BaselineEP	all	/	/	2,000	1,000	19.34	18.99
BaselineIT	/	/	2,000	2,000	1,000	20.77	21.55
IT+TERM	/	14,025	2,000	2,000	1,000	**21.89**	**22.73**
IT+EP1	1,000	/	2,000	2,000	1,000	**20.97**	*21.08
IT+EP10	10,000	/	2,000	2,000	1,000	**21.16**	21.66
IT+EP10+TERM	10,000	14,025	2,000	2,000	1,000	**22.20**	**22.16**

- TectoMT:
  - All above the baselines
  - Best approach: both (IT+EP10+TERM)

- PBSMT:
  - Above the baselines only those with added terminology (IT+TERM and IT+EP10+TERM)
  - Adding a small portion of out-of-domain corpus negatively influences (IT+EP1)

Results (Aspect)

Aspect	Mean		Median		Mode		Sign.	IAA
	TectoMT	PBSMT	TectoMT	PBSMT	TectoMT	PBSMT		
Fluency	**1.78**	1.74	**2**	1.5	2	2	0.054	0.52
Adequacy	**2.28**	2.24	2	2	2	2	**0.047**	0.55
Total	**2.27**	2.23	2	2	2	2	**0.048**	0.55

- TectoMT achieved significantly higher Adequacy score and Total score

- TectoMT achieved higher Mean and Median value for Fluency (not statistically significant difference)

Results (Errors)

Errors	Mean		Median		Mode		Sign.	IAA
	TectoMT	PBSMT	TectoMT	PBSMT	TectoMT	PBSMT		
Orthographic	1.15	**0.95**	1.25	**1**	1.5	**1**	**0.001**	0.50
Morphologic	0.97	**0.74**	1	**0.5**	1	**0**	**0.000**	0.54
Syntactic	1.31	**1.26**	1.5	1.5	1.5	1.5	**0.045**	0.49
Semantic	**1.37**	1.50	1.5	1.5	2	2	**0.009**	0.53

- Number of Orthographic, Morphologic, and Syntactic errors is significantly higher in TectoMT than in PBSMT system.

- Number of Semantic errors is significantly higher in PBSMT than in TectoMT system.

Sentence-wise Comparison

Comparison	Scores			Number of errors			
	Fluency	Adequacy	Total	Ortho.	Morpho.	Synt.	Sem.
TectoMT>PBSMT	47	55	55	69	81	58	98
TectoMT=PBSMT	117	96	96	96	77	85	102
TectoMT<PBSMT	36	49	49	35	42	57	60

- Sentences generated by TectoMT represent more fluent and adequate translations, but they also have greater number of errors.
- These results indicate one of the following:
  - Fluency and adequacy cannot be well captured by these types of errors.
  - The errors produced by the TectoMT system are not as severe as those produced by the PBSMT system.

Conclusions

- Adding in-domain bilingual terminology significantly improves the performance of both systems (TectoMT and PBSMT).

- Adding a combination of in-domain bilingual terminology and out-of-domain sentence pairs significantly improves the performance of both systems (TectoMT and PBSMT).

- Adding only some portion of out-of-domain sentence pairs only improves the performance of TectoMT system, while it either impairs or does not significantly change the performance of the PBSMT system.

Limitations

- We used only the basic domain-adaptation technique for the PBSMT system.
- We used no domain-adaptation techniques for the TectoMT system.

Thank you!

Contact:
Sanja.stajner@di.fc.ul.pt
Joao.rodrigues@di.fc.ul.pt
Luis.gomes@di.fc.ul.pt
Antonio.branco@di.fc.ul.pt