USC Viterbi
Department of Computer Science

USC Institute for Creative Technologies

Language Portability for Dialogue Systems: Translating a Question-Answering System from English into Tamil

Satheesh Ravi and Ron Artstein

Materials

Subset from *New Dimensions in Testimony*
Question-answering dialogue system

Classifier: pick best response from fixed answer set
Language model (LM): use question vocabulary
Cross-language model (CLM): use answer vocabulary

English results:

Tokenizer	Accuracy (%)	
	LM	CLM
Simple	89	82
NLTK	89	79

Portability

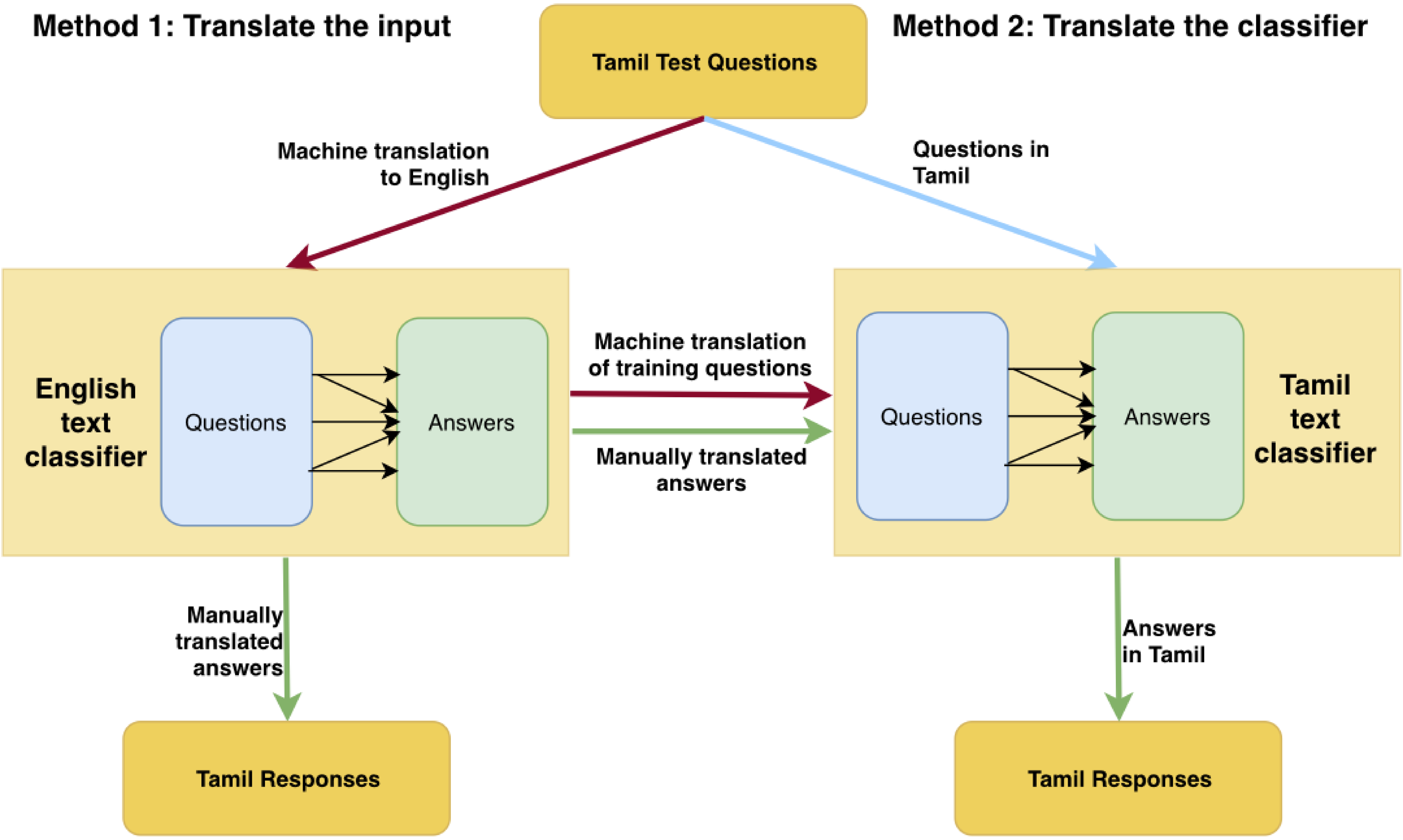
Test questions translated manually to Tamil to simulate actual use (28 in the experiment subset)

Responses translated manually to Tamil to ensure coherence (45 in the experiment subset)

Most of the work is creating training data: collecting questions (441 in experiment subset)
annotating links (1001 in experiment subset) ⇒ **Machine Translation**

Method 1: Translate the input

Method 2: Translate the classifier



Tamil Test Questions

Machine translation to English

Questions in Tamil

English text classifier

Questions

Answers

Machine translation of training questions

Manually translated answers

Questions

Answers

Tamil text classifier

Manually translated answers

Answers in Tamil

Tamil Responses

Tamil Responses

English Tokenizer	Accuracy (%)	
	LM	CLM
NLTK	79	57
Simple	64	46

⇐ Results ⇒

Question Translation	Accuracy (%)	
	LM	CLM
Manual	79	61
Machine	54	43

Discussion

Performance drop not too big in best case scenarios: **Promising approach**

Machine translation penalty: MT English worse than English
MT Tamil training data worse than Manually translated Tamil training data

Tamil penalty: Tamil classifier worse than English
English→Tamil MT worse than Tamil→English MT

Need better handling of Tamil morphology, e.g. case normalization:

எதிர்காலம்	எதிர்காலத்தின்
etirkaalam	etirkaalattin
'future' (unmarked)	'future' (genitive)