Discriminative Phrase Selection for Statistical Machine Translation

Cristina España-Bonet, Jesús Giménez and Lluís Màrquez

Universitat Politècnica de Catalunya

NIST'08 MT Evaluation Workshop

28th March, 2008

Overview

1. Introduction
2. Discriminative Phrase Selection
   - The method
   - Results for Arabic-to-English
3. Full Translation Task
   - The method
   - Results for Arabic-to-English
4. Conclusions

Statistical Machine Translation

$$\hat{e} = T(f) = \underset{e}{\operatorname{argmax}} P(e|f)$$

Statistical Machine Translation

$$\hat{e} = \operatorname*{argmax}_{e} \{\log \; P(e|f)\} = \operatorname*{argmax}_{e} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e, f) \right\}$$

Phrase-based Models, Log-linear extension...

Statistical Machine Translation

$$\hat{e} = \operatorname*{argmax}_{e} \{\log \; P(e|f)\} = \operatorname*{argmax}_{e} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e, f) \right\}$$

Phrase-based Models, Log-linear extension...

- but usually $P(e|f)$ is estimated by relative frequency counts (MLE)
- without context information (e.g., source-context is ignored)

Introduction
The general idea

Use Discriminative Machine Learning to estimate $P(e|f)$

Introduction

The general idea

Use Discriminative Machine Learning to estimate $P(e|f)$

- Vickrey et al. (2005)
- Carpuat and Wu (2006, 2007)
- Giménez and Màrquez (2007, 2008)
- Stroppa et al. (2007)
- Bangalore et al. (2007)

Use Discriminative Machine Learning to estimate $P(e|f)$

- Vickrey et al. (2005)
- Carpuat and Wu (2006, 2007)
- Giménez and Màrquez (2007, 2008)
- Stroppa et al. (2007)
- Bangalore et al. (2007)

Discriminative Phrase Selection:

☞ Phrase selection is treated as a classification problem

- We use SVMs to solve the multiclass classification problem
- Training set: phrase-aligned parallel corpus
- Every possible translation is a class → one-vs-all classification:

Discriminative Phrase Selection

The method

Discriminative Phrase Selection:

- Phrase selection is treated as a classification problem
- ☞ We use SVMs to solve the multiclass classification problem
- Training set: phrase-aligned parallel corpus
- Every possible translation is a class $\longrightarrow$ one-vs-all classification:

Discriminative Phrase Selection:

- Phrase selection is treated as a classification problem
- We use SVMs to solve the multiclass classification problem

☞ Training set: phrase-aligned parallel corpus

- Every possible translation is a class $\longrightarrow$ one-vs-all classification:

Discriminative Phrase Selection:

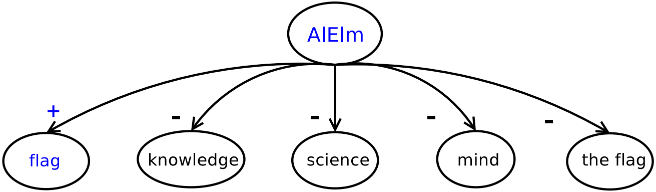
- Phrase selection is treated as a classification problem
- We use SVMs to solve the multiclass classification problem
- Training set: phrase-aligned parallel corpus

☞ Every possible translation is a class $\longrightarrow$ one-vs-all classification:

Discriminative Phrase Selection

The method

`wAn$d AllbnAnywn Al*yn HmlwA ktb SlAp w rfEwA AlElm AllbnAny , Aln$yd AlwTny AllbnAny .`

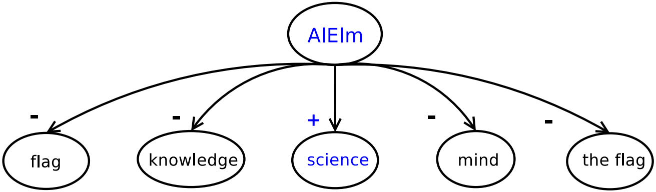


`The Lebanese , who came carrying prayer books and the Lebanese flag , sang the Lebanese national anthem .`

Discriminative Phrase Selection

The method

>n HAlp AlElm w AltknwlwjyA ldY nA fy nhAyp Alqrn AlE$ryn l hA ElAmtAn mhmtAn . Al>wly gyAb AlmlAHqp fy h*A AlqTAE .



The situation of science and technology in Egypt at the end of the 20th century had two important features .

Discriminative Phrase Selection

The method

SVMs allow to use context and linguistic information

Features set for the SVMs include:

- Source phrase features
  - PoS, coarse PoS and chunk *n*-grams
- Source sentence features
  - Word, PoS, coarse PoS, chunk *n*-grams and bag-of-words

The training set

News domain:

- 123,662 lines, 3.9M Arabic tokens, 4.2M English tokens

Corpus	Lines	Arabic tokens	English tokens
Arabic English Parallel News Part 1	61,000	2,179,289	2,273,021
Arabic News Translation Text Part 1	18,000	532,771	602,262
Arabic Treebank English Translation	23,800	660,821	739,695
eTIRR Arabic English News Text	4,000	97,882	98,655
Multiple-Translation Arabic	15,533	434,465	507,617
TIDES MT2004 Arabic evaluation data	1,329	40,667	47,324
Total:	123,662	3,945,895	4,262,740

Phrase translation task

Improvement in accuracy wrt. the most frequent translation, MFT

Training set occurrences	#	Acc.MFT (%)	Acc.DPT (%)
100-500	4,310	58.7	66.5
501-1,000	565	62.3	68.8
1,001-5,000	393	66.7	73.0
5,001-10,000	27	72.2	79.5
10,001-50,000	19	66.6	74.8
$>$ 50,000	7	76.2	80.7
Total:	5,321	59.8	67.3

Full Translation Task

Integration into a SMT system

Estimation of the discriminative phrase translation model and integration into the SMT system:

- ☞ Training linear SVMs (SVM^{light}, Joachims 1999) for every translation of every phrase
  - Convert SVM score into probability via a *softmax* function
  - Include this probability in the translation model within a Log-linear model

Full Translation Task

Integration into a SMT system

Estimation of the discriminative phrase translation model and integration into the SMT system:

- Training linear SVMs (SVM^{light}, Joachims 1999) for every translation of every phrase
- ☞ Convert SVM score into probability via a *softmax* function
- Include this probability in the translation model within a Log-linear model

Estimation of the discriminative phrase translation model and integration into the SMT system:

- Training linear SVMs (SVM^{light}, Joachims 1999) for every translation of every phrase
- Convert SVM score into probability via a *softmax* function

☞ Include this probability in the translation model within a Log-linear model

Full Translation Task

Integration into the SMT system

$\texttt{Hyv}_{28}$ $\texttt{tm}_{22}$ $\texttt{AHrAq}$ $\texttt{AlElm}_1$ $\texttt{AldnmArky}$ $\cdot_{1128}$

Translation table example:

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(e\|f)$
AlElm_1	flag	0.1986	0.3241
AlElm_1	the	0.0419	0.0207
AlElm_1	mind	0.0401	0.0620
AlElm_1	the flag	0.0397	0.0414
AlElm_1	flag during	0.0394	0.0138
AlElm_1	knowledge	0.0392	0.1103
AlElm_1	flag caused	0.0387	0.0138
AlElm_1	science	0.0377	0.1793
AlElm_1	education	0.0377	0.0138
AlElm_1	in mind	0.0371	0.0138
...			

Full Translation Task

Integration into the SMT system

$\texttt{Hyv}_{28}$ $\texttt{tm}_{22}$ $\texttt{AHrAq}$ $\texttt{AlElm}_1$ $\texttt{AldnmArky}$ $._{1128}$

Translation table example:

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(e\|f)$
AlElm_1	flag	0.1986	0.3241
AlElm_1	the	0.0419	0.0207
AlElm_1	mind	0.0401	0.0620
AlElm_1	the flag	0.0397	0.0414
AlElm_1	flag during	0.0394	0.0138
AlElm_1	knowledge	0.0392	0.1103
AlElm_1	flag caused	0.0387	0.0138
AlElm_1	science	0.0377	0.1793
AlElm_1	education	0.0377	0.0138
AlElm_1	in mind	0.0371	0.0138
...			

Full Translation Task

Results for Arabic-to-English

Building the system:

- Language model
  - 5-gram Language Model, interpolated Kneser-Ney discounting
  - SRILM Toolkit (Stolcke 2002)
- Translation model
  - Alignments: GIZA++ Toolkit (Och & Ney 2003)
  - Translation tables: Moses package (Koehn et al. 2006)
    MLT package (Giménez & Màrquez)
- Decoder
  - Moses decoder (Koehn et al. 2006)

Full Translation Task

Results for Arabic-to-English

Building the system:

- Language model
  - 5-gram Language Model, interpolated Kneser-Ney discounting
  - SRILM Toolkit (Stolcke 2002)
- ☞ Translation model
  - ☞ Alignments: GIZA++ Toolkit (Och & Ney 2003)
  - ☞ Translation tables: Moses package (Koehn et al. 2006)
    MLT package (Giménez & Màrquez)
- Decoder
  - Moses decoder (Koehn et al. 2006)

Full Translation Task

Results for Arabic-to-English

Building the system:

- Language model
  - 5-gram Language Model, interpolated Kneser-Ney discounting
  - SRILM Toolkit (Stolcke 2002)
- Translation model
  - Alignments: GIZA++ Toolkit (Och & Ney 2003)
  - Translation tables: Moses package (Koehn et al. 2006)
    MLT package (Giménez & Màrquez)
- ☞ Decoder
  - ☞ Moses decoder (Koehn et al. 2006)

Full Translation Task

Results for Arabic-to-English

Translation table (input for the Moses decoder)

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(f\|e)$	$lex(f\|e)$	$P_{MLE}(e\|f)$	$lex(e\|f)$
$AlElm_1$	flag	0.1986	0.6438	0.5417	0.3241	0.2826
$AlElm_1$	science	0.0377	0.1529	0.1477	0.1793	0.1413
...						

Full Translation Task

Results for Arabic-to-English

Translation table (input for the Moses decoder)

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(f\|e)$	$lex(f\|e)$	$P_{MLE}(e\|f)$	$lex(e\|f)$
$AlElm_1$	flag	0.1986	0.6438	0.5417	0.3241	0.2826
$AlElm_1$	science	0.0377	0.1529	0.1477	0.1793	0.1413
...						

Three systems:

SMT
standard

Full Translation Task

Results for Arabic-to-English

Translation table (input for the Moses decoder)

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(f\|e)$	$lex(f\|e)$	$P_{MLE}(e\|f)$	$lex(e\|f)$
AlElm_1	flag	0.1986	0.6438	0.5417	0.3241	0.2826
AlElm_1	science	0.0377	0.1529	0.1477	0.1793	0.1413
...						

Three systems:

SMT	DPT
standard	replace MLE

Full Translation Task

Results for Arabic-to-English

Translation table (input for the Moses decoder)

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(f\|e)$	$lex(f\|e)$	$P_{MLE}(e\|f)$	$lex(e\|f)$
AlElm$_1$	flag	0.1986	0.6438	0.5417	0.3241	0.2826
AlElm$_1$	science	0.0377	0.1529	0.1477	0.1793	0.1413
...						

Three systems:

SMT	DPT	DPT$^+$
standard	replace MLE	add to MLE

Full Translation Task

Results for Arabic-to-English

Translation table (input for the Moses decoder)

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(f\|e)$	$lex(f\|e)$	$P_{MLE}(e\|f)$	$lex(e\|f)$
AlElm$_1$	flag	0.1986	0.6438	0.5417	0.3241	0.2826
AlElm$_1$	science	0.0377	0.1529	0.1477	0.1793	0.1413
...						

Three systems:

SMT	DPT	DPT$^+$
standard	replace MLE	add to MLE

Automatic evaluation: $\mathrm{IQ}_{\mathrm{MT}}$ package (yesterday's talk!)

Full Translation Task

Results for Arabic-to-English

Lexical metrics

Best system:

DPT

	SMT	DPT	DPT$^+$
1-PER	0.5814	0.5892	0.5852
1-TER	0.4493	0.4482	0.4454
1-WER	0.4161	0.4102	0.4078
BLEU-4	0.3103	0.3243	0.3175
NIST-5	8.7113	8.9053	8.7920
GTM-1	0.6974	0.7159	0.7107
GTM-2	0.2234	0.2267	0.2247
GTM-3	0.1721	0.1745	0.1728
RG-L	0.4986	0.4993	0.4968
RG-S*	0.3185	0.3229	0.3188
RG-SU*	0.3395	0.3437	0.3395
RG-W-1.2	0.2662	0.2675	0.2659
MTR-exact	0.4909	0.5001	0.4958
MTR-stem	0.5098	0.5174	0.5135
MTR-wnstm	0.5147	0.5222	0.5186
MTR-wnsyn	0.5352	0.5426	0.5391

Full Translation Task

Results for Arabic-to-English

Lexical metrics

Best system:

DPT

+1.4 BLEU improvement

	SMT	DPT	DPT$^+$
1-PER	0.5814	0.5892	0.5852
1-TER	0.4493	0.4482	0.4454
1-WER	0.4161	0.4102	0.4078
BLEU-4	0.3103	0.3243	0.3175
NIST-5	8.7113	8.9053	8.7920
GTM-1	0.6974	0.7159	0.7107
GTM-2	0.2234	0.2267	0.2247
GTM-3	0.1721	0.1745	0.1728
RG-L	0.4986	0.4993	0.4968
RG-S*	0.3185	0.3229	0.3188
RG-SU*	0.3395	0.3437	0.3395
RG-W-1.2	0.2662	0.2675	0.2659
MTR-exact	0.4909	0.5001	0.4958
MTR-stem	0.5098	0.5174	0.5135
MTR-wnstm	0.5147	0.5222	0.5186
MTR-wnsyn	0.5352	0.5426	0.5391

Syntactic metrics

Best system:

DPT

		SMT	DPT	DPT^{+}
Shallow	SP-Oc-*	0.4376	0.4448	0.4407
	SP-Op-*	0.4195	0.4271	0.4235
	SP-cNIST-5	5.5783	5.6684	5.6703
	SP-iobNIST-5	5.9931	6.1318	6.1172
	SP-lNIST-5	8.8869	9.0547	8.9523
	SP-pNIST-5	6.9679	7.1610	7.1117
Constituent Parsing	CP-Oc-*	0.3943	0.3995	0.3962
	CP-Op-*	0.4220	0.4296	0.4265
	CP-STM-9	0.2396	0.2394	0.2380
Dependency Parsing	DP-Oc-*	0.3852	0.3949	0.3892
	DP-Ol-*	0.3051	0.3164	0.3115
	DP-Or-*	0.2523	0.2557	0.2534
	DP-HWC-c-4	0.2986	0.2975	0.2970
	DP-HWC-r-4	0.2023	0.2023	0.2029
	DP-HWC-w-4	0.0835	0.0826	0.0831

Full Translation Task

Results for Arabic-to-English

Semantic metrics

Best system:

no clear *winner*, but slight advantage in favor of DPT/DPT$^+$

		SMT	DPT	DPT$^+$
Semantic Role	SR-Mr-*	0.0224	0.0227	0.0262
	SR-Mrv-*	0.0123	0.0129	0.0129
	SR-Or	0.3686	0.3792	0.3609
	SR-Or-*	0.1160	0.1209	0.1234
	SR-Orv	0.0685	0.0815	0.0765
	SR-Orv-*	0.0284	0.0325	0.0349
Discourse Representation	DR-Or-*	0.2121	0.2115	0.2094
	DR-Orp-*	0.3296	0.3252	0.3245
	DR-STM-9	0.1787	0.1902	0.1872

Conclusions

- Improvements at a lexical and syntactic level with respect to our baseline with DPT system
- Further improvements expected with a better integration into the SMT system
- Local accuracy improvement is not fully reflected in global MT quality:
  1. language model
  2. local classifiers are not trained in the context of the global task

Conclusions

- Improvements at a lexical and syntactic level with respect to our baseline with DPT system

☞ Further improvements expected with a better integration into the SMT system

- Local accuracy improvement is not fully reflected in global MT quality:
  1. language model
  2. local classifiers are not trained in the context of the global task

Conclusions

- Improvements at a lexical and syntactic level with respect to our baseline with DPT system

- Further improvements expected with a better integration into the SMT system

☞ Local accuracy improvement is not fully reflected in global MT quality:

1. language model
2. local classifiers are not trained in the context of the global task

Thank you!

Differences between ML techniques to estimate $P(e|f)$

Related work, approaches

Task Differences

- Language pair
- Domain

System Differences

- System Architecture
- Learning scheme

Evaluation Differences

- Metrics
- Manual evaluations

Differences between ML techniques to estimate $P(e|f)$

Related work, approaches

Task Differences

- Language pair
  - Spanish-to-English
  - Chinese-to-English
  - Arabic-to-English
  - French-to-English
- Domain
  - Europarl
  - NIST
  - BTEC
  - Hansards
  - United Nations

Differences between ML techniques to estimate $P(e|f)$

Related work, approaches

System Differences

- System Architecture
  - Log-linear models
  - Finite-state transducers (lexical selection + sentence reconstruction)
- Learning scheme
  - SVMs
  - Maximum entropy
  - Combination of maximum entropy, naïve Bayes, boosting, Kernel PCA-based models
  - Memory-based learning

Differences between ML techniques to estimate $P(e|f)$

Related work, approaches

Evaluation Differences

- Metrics
  - bleu
  - + nist
  - + lexical similarity
  - + syntactic and semantic similarities
  - + metric combinations
- Manual evaluations

Local Phrase Translation Accuracy
Evaluation Schemes

- Different settings depending on the number of examples available

#examples	evaluation scheme: development and test	evaluation scheme: test only
$2 - 9$	leave-one-out	
$10..99$	10-fold cross validation	
$100..499$	5-fold cross validation	
$500..999$	3-fold cross validation	
$1,000..4,999$	train(80%)–dev(10%)–test(10%)	train(90%)–test(10%)
$5,000..9,999$	train(70%)–dev(15%)–test(15%)	train(80%)–test(20%)
$> 10,000$	train(60%)–dev(20%)–test(20%)	train(75%)–test(25%)

- Automatic phrase-alignments as a gold-standard

Discriminative Phrase Selection for Statistical Machine Translation

Cristina España-Bonet, Jesús Giménez and Lluís Màrquez

Universitat Politècnica de Catalunya

NIST'08 MT Evaluation Workshop

28th March, 2008