Transductive learning for statistical machine translation

Nicola Ueffing[1] Gholamreza Haffari[2] Anoop Sarkar[2]

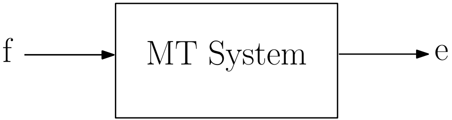
[1]Interactive Language Technologies Group
National Research Council Canada
Gatineau, QC, Canada
nicola.ueffing@nrc.gc.ca

[2]School of Computing Science
Simon Fraser University
Vancouver, Canada
{ghaffar1,anoop}@cs.sfu.ca

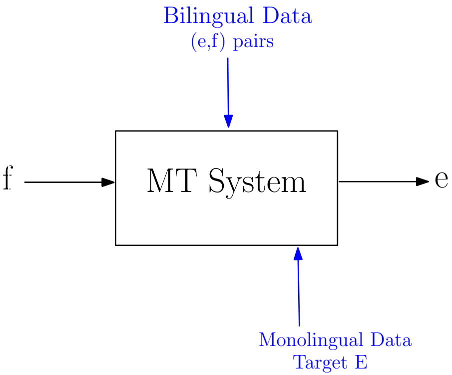
ACL 2007: June 25

Outline

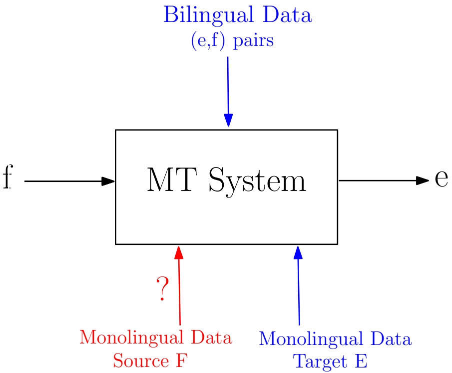
Motivation



Motivation


Bilingual Data
(e,f) pairs
f
MT System
e
Monolingual Data
Target E

Motivation



Here: we explore monolingual source-language data to improve translation quality

Where it would be useful?

- In some cases amount of bilingual data is limited and expensive to create
- Use monolingual source-language data to
  - adapt to new domain, topic or style
  - overcome training/testing data mismatch, e.g. text/speech

Where it would be useful?

- In some cases amount of bilingual data is limited and expensive to create
- Use monolingual source-language data to
  - adapt to new domain, topic or style
  - overcome training/testing data mismatch, e.g. text/speech

Examples:

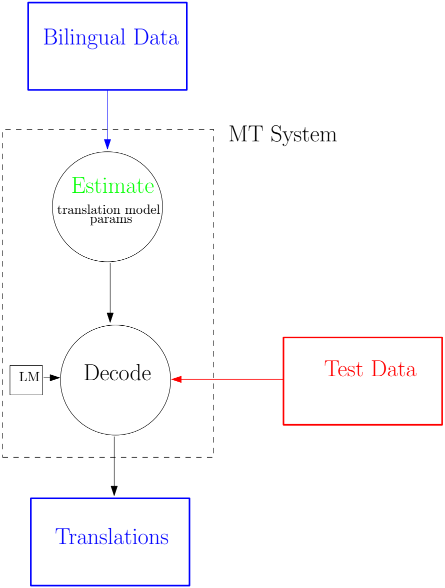
training data	testing data	effect
newswire	web text	adapt to domain and style
written text	speech	adapt to speech characteristics
written text and speech	speech	identify parts of model relevant for speech

1 Motivation

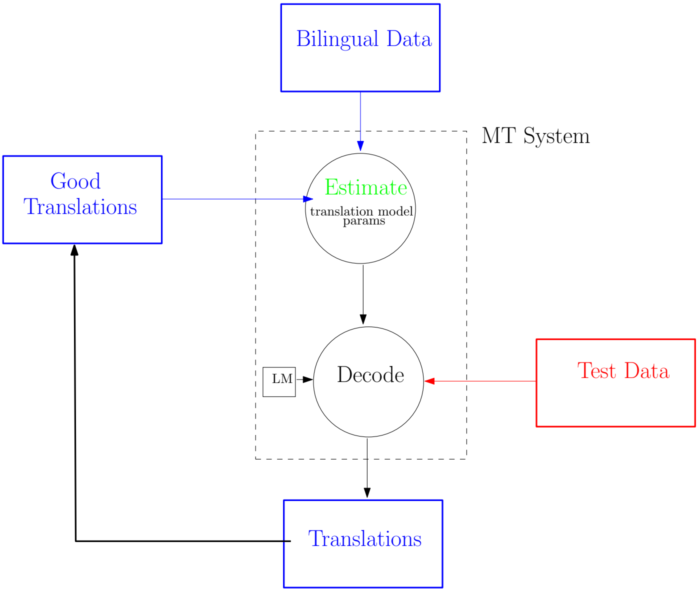
2 Transductive Machine Translation

3 Experimental Results
- SMT System
- EuroParl French–English
- NIST Chinese–English

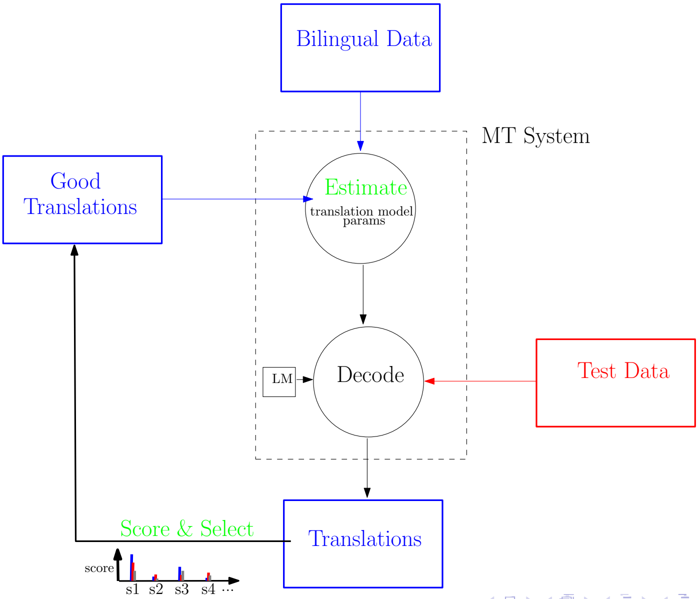
Transductive SMT



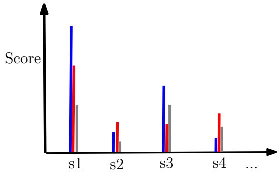
Transductive SMT



Transductive SMT

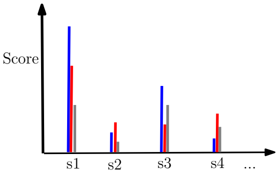


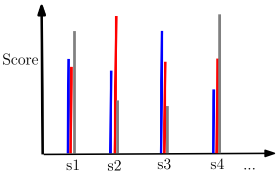
Scoring Translations



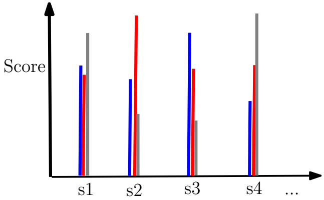
1. Confidence estimation
   - log-linear combination of different posterior probabilities and LM probability
   - posterior probabilities for words and phrases, calculated over N-best list
   - combination optimized w.r.t. sentence classification error rate

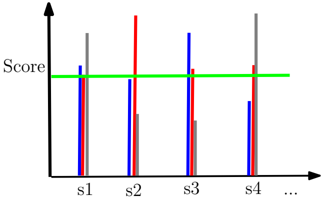
Scoring Translations





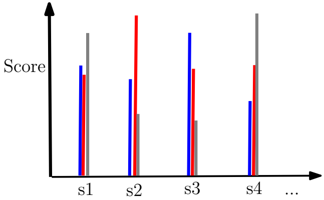
1. Confidence estimation
   - log-linear combination of different posterior probabilities and LM probability
   - posterior probabilities for words and phrases, calculated over N-best list
   - combination optimized w.r.t. sentence classification error rate
2. Normalized sentence score assigned by SMT system

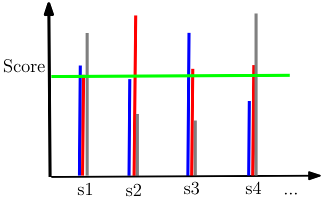




1. Importance sampling: sample with replacement, probability distribution based on scores

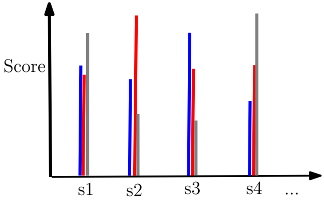
Selection

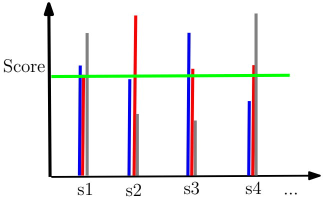




1. Importance sampling: sample with replacement, probability distribution based on scores
2. Threshold: select all translations with score above threshold, optimize threshold on dev set beforehand

Selection





1. Importance sampling: sample with replacement, probability distribution based on scores
2. Threshold: select all translations with score above threshold, optimize threshold on dev set beforehand
3. Keep *all* translations: comparative experiment

Estimate

- We extract "good" translations and use these to augment our SMT system

Estimate

- We extract “good” translations and use these to augment our SMT system
- Different choices are used to Estimate a new model

Estimate

- We extract "good" translations and use these to augment our SMT system
- Different choices are used to Estimate a new model

1. Add new translations to training set and do full re-training (can be made efficient; details in the paper)

Estimate

- We extract "good" translations and use these to augment our SMT system
- Different choices are used to Estimate a new model

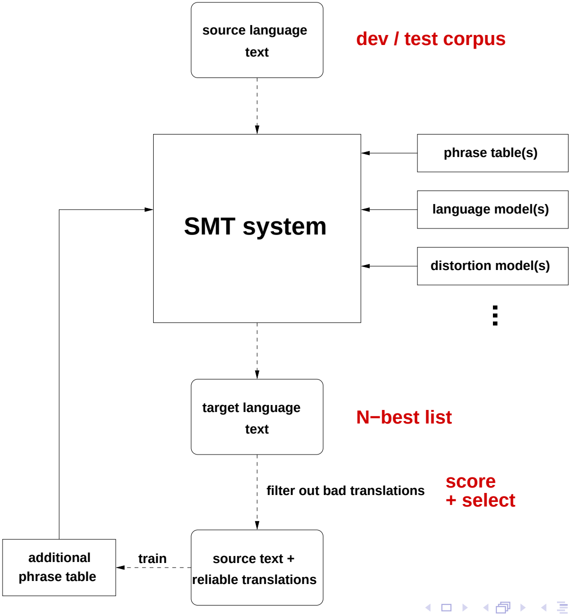
1. Add new translations to training set and do full re-training (can be made efficient; details in the paper)
2. A mixture model of phrase pair probabilities from training set combined with phrase pairs from dev/test set

Estimate

- We extract "good" translations and use these to augment our SMT system
- Different choices are used to Estimate a new model

1. Add new translations to training set and do full re-training (can be made efficient; details in the paper)
2. A mixture model of phrase pair probabilities from training set combined with phrase pairs from dev/test set
3. Use new phrase pairs to train an additional phrase table and use it as a new feature function in the SMT log-linear model (feature weights learned using dev corpus).

Estimate (additional phrase table)



Why does it work?

- Reinforces parts of the phrase translation model which are relevant for test corpus,
  obtain more focused probability distribution
- Composes new phrases, for example:

original parallel corpus	additional source data	possible new phrases
'A B', 'C D E'	'A B C D E'	'A B C', 'B C D E', 'A B C D E', ...

Limitations of the approach

- No learning of translations of *unknown* source-language words occurring in the new data
- Only learning of *compositional* phrases; system will not learn translation of idioms:

"it is raining"	+	"cats and dogs"	$\rightarrow$	"it is raining cats and dogs"
"es regnet"	+	"Katzen und Hunde"	$\not\rightarrow$	"es regnet in Strömen"
"il pleut"	+	"des chats et des chiens"	$\not\rightarrow$	"il pleut des cordes"

1 Motivation

2 Transductive Machine Translation

3 Experimental Results
- SMT System
- EuroParl French–English
- NIST Chinese–English

Experimental setting: Baseline & SMT system

PORTAGE: state-of-the-art phrase-based system (NRC, Canada)

Decoder models:

- several (smoothed) phrase table(s), translation direction $p(s_1^J \mid t_1^I)$
- several 4-gram language model(s), trained with SRILM toolkit
- distortion penalty based on number of skipped source words
- word penalty

Experimental setting: Baseline & SMT system

PORTAGE: state-of-the-art phrase-based system (NRC, Canada)

Decoder models:

- several (smoothed) phrase table(s), translation direction $p(s_1^J \mid t_1^I)$
- several 4-gram language model(s), trained with SRILM toolkit
- distortion penalty based on number of skipped source words
- word penalty

Additional rescoring models:

- two different IBM-1 features in both translation directions
- posterior probabilities for words, phrases, n-grams, and sentence length: calculated over the N-best list, using the sentence probabilities assigned by the baseline system

Our approach also works with other phrase-based MT system, e.g. Moses

EuroParl French–English

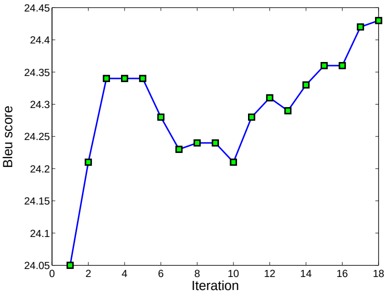
Setup and evaluation:

- French $\rightarrow$ English translation
- training and testing conditions: WMT2006 shared task
  688k sentence pairs for training
  2,000/3,064 sentences in dev/test set
- evaluate with BLEU-4, mWER, mPER, using 1 references
- 95%-confidence intervals, using bootstrap resampling

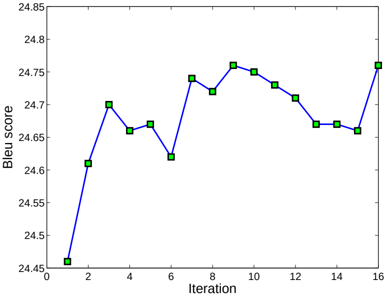
Results EuroParl French–English

Translation quality for importance sampling based on normalized sentence scores, full re-training of phrase table

Train100k



Train150k



Transductive learning provides improvement in accuracy equivalent to adding 50k training examples

EuroParl translation examples

baseline	but it will be agreed on what we are putting into this constitution .
adapted	but it must be agreed upon what we are putting into the constitution .
reference	but we must reach agreement on what to put in this constitution .
baseline	this does not want to say first of all , as a result .
adapted	it does not mean that everything is going on .
reference	this does not mean that everything has to happen at once .

NIST Chinese–English

Setup and evaluation:

- Chinese → English translation
- training conditions: NIST 2006 eval, large data track
- testing: 2006 eval corpus with 3,940 sentences
  4 different genres, partially not covered by training data
  (broadcast conversations, . . .)
- evaluate with BLEU-4, mWER, mPER, using 4 / 1 references
- 95%-confidence intervals, using bootstrap resampling

Results: NIST Chinese–English

Translation quality on NIST 2006 Chinese–English, NIST part.
Different versions of selection and scoring method.

selection	scoring	BLEU[%]	mWER[%]	mPER[%]
baseline		27.9±0.7	67.2±0.6	44.0±0.5

Results: NIST Chinese–English

Translation quality on NIST 2006 Chinese–English, NIST part.
Different versions of selection and scoring method.

selection	scoring	BLEU[%]	mWER[%]	mPER[%]
baseline		27.9±0.7	67.2±0.6	44.0±0.5
keep all		28.1	66.5	44.2

Translation quality on NIST 2006 Chinese–English, NIST part. Different versions of selection and scoring method.

selection	scoring	BLEU[%]	mWER[%]	mPER[%]
baseline		27.9±0.7	67.2±0.6	44.0±0.5
keep all		28.1	66.5	44.2
import. sampl.	norm.score	28.7	66.1	43.6
	confidence	28.4	65.8	43.2

Results: NIST Chinese–English

Translation quality on NIST 2006 Chinese–English, NIST part.
Different versions of selection and scoring method.

selection	scoring	BLEU[%]	mWER[%]	mPER[%]
baseline		27.9±0.7	67.2±0.6	44.0±0.5
keep all		28.1	66.5	44.2
import. sampl.	norm.score	28.7	66.1	43.6
	confidence	28.4	65.8	43.2
threshold	norm.score	28.3	66.1	43.5
	confidence	29.3	65.6	43.2

NIST translation examples

baseline	[the report said] [that the] [united states] [is] [a potential] [problem] [, the] [practice of] [china 's] [foreign policy] [is] [likely to] [weaken us] [influence] [.]
transductive	[the report] [said that] [this is] [a potential] [problem] [in] [the united states] [,] [china] [is] [likely to] [weaken] [the impact of] [american foreign policy] [.]
reference	the report said that this is a potential problem for america . china 's course of action could possibly weaken the influence of american foreign policy .

NIST translation examples

baseline	[the report said] [that the] [united states] [is] [a potential] [problem] [, the] [practice of] [china 's] [foreign policy] [is] [likely to] [weaken us] [influence] [.]
transductive	[the report] [said that] [this is] [a potential] [problem] [in] [the united states] [,] [china] [is] [likely to] [weaken] [the impact of] [american foreign policy] [.]
reference	the report said that this is a potential problem for america . china 's course of action could possibly weaken the influence of american foreign policy .
baseline	[what we advocate] [his] [name]
transductive	[we] [advocate] [him] [.]
reference	we advocate him .

NIST translation examples

baseline	[the report said] [that the] [united states] [is] [a potential] [problem] [, the] [practice of] [china 's] [foreign policy] [is] [likely to] [weaken us] [influence] [.]
transductive	[the report] [said that] [this is] [a potential] [problem] [in] [the united states] [,] [china] [is] [likely to] [weaken] [the impact of] [american foreign policy] [.]
reference	the report said that this is a potential problem for america . china 's course of action could possibly weaken the influence of american foreign policy .
baseline	[what we advocate] [his] [name]
transductive	[we] [advocate] [him] [.]
reference	we advocate him .
baseline	["] [we should] [really be] [male] [nominees] [..] [....]
transductive	[he] [should] [be] [nominated] [male] [,] [really] [.]
reference	he should be nominated as the best actor , really .

Conclusion

- Explore monolingual source-language data to improve an existing MT system:
    - translate data using MT system
    - automatically identify reliable translations
    - learn new models on these

Conclusion

- Explore monolingual source-language data to improve an existing MT system:
  - translate data using MT system
  - automatically identify reliable translations
  - learn new models on these
- Introduced transductive learning approach for statistical MT
  - filtering training data for re-training
  - using additional phrase table from test data as feature in MT log-linear model
  - confidence estimation for accurate detection of good translations
  - importance sampling with thresholding to obtain multiple good translations even for a single sentence

Conclusion

- Explore monolingual source-language data to improve an existing MT system:
  - translate data using MT system
  - automatically identify reliable translations
  - learn new models on these
- Introduced transductive learning approach for statistical MT
  - filtering training data for re-training
  - using additional phrase table from test data as feature in MT log-linear model
  - confidence estimation for accurate detection of good translations
  - importance sampling with thresholding to obtain multiple good translations even for a single sentence
- Translation quality improves through transductive learning

Conclusion

- Explore monolingual source-language data to improve an existing MT system:
  - translate data using MT system
  - automatically identify reliable translations
  - learn new models on these
- Introduced transductive learning approach for statistical MT
  - filtering training data for re-training
  - using additional phrase table from test data as feature in MT log-linear model
  - confidence estimation for accurate detection of good translations
  - importance sampling with thresholding to obtain multiple good translations even for a single sentence
- Translation quality improves through transductive learning
- Discarding bad translations is important

Conclusion

- Explore monolingual source-language data to improve an existing MT system:
  - translate data using MT system
  - automatically identify reliable translations
  - learn new models on these
- Introduced transductive learning approach for statistical MT
  - filtering training data for re-training
  - using additional phrase table from test data as feature in MT log-linear model
  - confidence estimation for accurate detection of good translations
  - importance sampling with thresholding to obtain multiple good translations even for a single sentence
- Translation quality improves through transductive learning
- Discarding bad translations is important
- Approach applicable to other types of statistical MT system

Literature

- Transductive learning/unsupervised training: D. Yarowsky [ACL, 1995], Abney [CompLing 30-03, 2004], Vapnik "Statistical learning theory" [Wiley, 1998]
- Self-training for SMT: Ueffing [IWSLT, 2006]
- PORTAGE: Ueffing et. al. [ACL WMT Workshop, 2007]
- Confidence measures: Blatz et al. [CoLing 2004], Ueffing and Ney [CompLing 33-01, 2007]

Acknowledgment

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

This research was partially supported by NSERC, Canada (RGPIN: 264905).

END

Filtering the training corpus

- If the size of the training corpus is huge, the training time is going to be very long;
- filter training corpus based on n-gram-coverage with the dev/test corpus to find relevant parts

Statistics of the phrase tables trained on the genres of the NIST test corpora.

Chinese–English eval-04	editorials	newswire	speeches	
sentences	449	901	438	
selected translations	101	187	113	
size of adapted phrase table	1,981	3,591	2,321	
new phrases in phrase table	679	1,359	657	
adapted phrases used	707	1,314	815	
new phrases used	23	47	25	
Chinese–English eval-06	broadcast conversations	broadcast news	newsgroup	newswire
sentences	979	1,083	898	980
selected translations	477	274	226	172
size of adapted phrase table	2,155	4,027	2,905	2,804
new phrases in phrase table	1,058	1,645	1,259	1,058
adapted phrases used	759	1,479	1,077	1,115
new phrases used	90	86	88	41

Results: NIST Chinese–English

Translation quality on the NIST 2006 Chinese–English task.
Different versions of selection and scoring method.

corpus	selection	scoring	BLEU[%]	mWER[%]	mPER[%]
GALE	baseline		12.7±0.5	75.8±0.6	54.6±0.6
(1 ref.)	keep all		12.9	75.7	55.0
	import.sampl.	norm.score	13.2	74.7	54.1
		confidence	12.9	74.4	53.5
	threshold	norm.score	12.7	75.2	54.2
		confidence	13.6	73.4	53.2
NIST	baseline		27.9±0.7	67.2±0.6	44.0±0.5
(4 refs.)	keep all		28.1	66.5	44.2
	import.sampl.	norm.score	28.7	66.1	43.6
		confidence	28.4	65.8	43.2
	threshold	norm.score	28.3	66.1	43.5
		confidence	29.3	65.6	43.2

More NIST translation examples (1)

baseline	[the capitalist] [system] [, because] [it] [is] [immoral] [to] [criticize] [china] [for years] [, capitalism] [, so] [it] [didn't] [have] [a set of] [moral values] [.]
transductive	[capitalism] [has] [a set] [of] [moral values] [,] [because] [china] [has] [denounced] [capitalism] [,] [so it] [does not] [have] [a set] [of moral] [.]
reference	capitalism , its set of morals , because china has criticized capitalism for many years , this set of morals is no longer there .
baseline	[the fact] [that this] [is] [.]
transductive	[this] [is] [the point] [.]
reference	that is actually the point .

Results EuroParl French–English

Translation quality for importance sampling with full re-training, normalized sentence scores, filtered 100k training sentence pairs