

Improved Statistical MT

Chris Callison-Burch
University of Edinburgh &
LINEAR B LTD.

Overview

- Review of statistical translation
- Phrase-based statistical translation
- How editing can improve SMT

Advantages of SMT

- Can be applied to any language pair
- Quick to develop
- Improves as more data becomes available
- (Recently) High-quality

Overview of SMT

- Goal: find sentence which is most likely translation of sentence, $p(e|f)$
- $e^* = \operatorname{argmax} p(e) p(f|e)$
- $p(e)$ is language model
- $p(f|e)$ is translation model

Estimating probabilities

- Direct estimation is not possible
- So decompose problem into smaller units
- Brown *et al.* use word-level alignments:

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

Word-level alignments

	France	was	ordered	to	bear	its	own	costs
La	■							
République	■							
française	■							
supportera		■	■	■	■			
ses						■		
propres							■	
dépens								■

Translation probs from alignment probs

- $p(f|e)$ is larger for sentence pairs with probable alignments
- So $p(\text{La République française supportera ses propres dépens} \mid \text{France was ordered to bear its own costs}) > p(\text{La République française supportera ses propres dépens} \mid \text{I like cheese})$

IBM Models

- Strange parameters
 - Fertility
 - Translation
 - Spurious Word
 - Distortion
- String rewriting

Problems

- Fertility poor representation of multi-word translations
- Lots of re-ordering has to happen
- *These are solved by phrase-based translation*

Example Source

- Honorables sénateurs, tandis que la guerre en Irak entre dans sa troisième semaine, nous ne devons pas oublier qu'il faut prendre des mesures pour éviter une crise humanitaire dans la population civile. À cet égard, il y a de nombreux domaines dans lesquels les Canadiens doivent faire preuve de leadership.

Maintenant que la guerre fait rage en Irak et que des pénuries de produits alimentaires et de fournitures médicales commencent à se produire, les pays qui ont accès à des ressources ont la responsabilité d'essayer de minimiser les effets négatifs du conflit sur la population irakienne. Je crois personnellement que le Canada devrait jouer un plus grand rôle à cet égard.

Au Canada, nous disposons en abondance de blé et d'autres produits alimentaires. Nous pouvons également fournir des articles médicaux aux citoyens de l'Irak. Le Canada a une fière réputation pour ce qui est d'offrir une aide humanitaire aux gens quand ils en ont besoin.

Word-based SMT

- Honourable senators, while that the war in Iraq between in his third week, we not must not forget that it must take of measures to avoid a crisis humanitarian in the people calendar. In many areas which Canadians must show leadership in this regard

Now that the war fact raging in Iraq and that of shortages of products food and of supplies medical beginning to be produce, the country which have access to of resources have the responsibility of try of minimize the effects negative of conflict on the people irakienne. I personally believe this regard great role Canada should play more

In other food products wheat Canada have abundant We can provide the citizens medical articles Iraq The offer Canada has a reputation proud

Phrase-based SMT

- Honourable senators, while the war in Iraq extending into a third week, the minister should not forget the one we must take some steps to prevent a crisis humanitarian face in the civilian populations. In this regard, there are a number of areas in which the Canadians must concentrate to show leadership.

That the war is raging in Iraq and that a shortage of food and medical supplies are beginning to take place, those countries that have access to the resources are the responsibility for doing try to minimize the negative effects on workers on the people irakienne. I personally believe that Canada should play a more significant role in this regard.

In Canada, we have in abundance of wheat, as have other food products. We can also provision of medical articles to the people on the other Iraq. Canada has a proud record in so far, as would be to give humanitarian assistance to people when they need it.

Larger Units

Je crois personnellement que le	Canada devrait jouer	un plus grand rôle	à cet égard.
---------------------------------	----------------------	--------------------	--------------

I personally believe that	Canada should play	a more significant role	in this regard.
---------------------------	--------------------	-------------------------	-----------------

Phrase Probabilities

- Simple to calculate

- $$p(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f}, \bar{e})}$$

- Trick is to figure out how to collect counts, which phrases pair up

	France	was	ordered	to	bear	its	own	costs
La	■							
République	■							
française	■							
supportera		■	■	■	■			
ses						■		
propres							■	
dépens								■

France, La République française
 was ordered to bear, supportera
 its, ses
 own, propres
 costs, dépens

	France	was	ordered	to	bear	its	own	costs
La								
République								
française								
supportera								
ses								
propres								
dépens								

France was ordered to bear, La République française supportera
 was ordered to bear its, supportera ses
 its own, ses propres
 own costs, propres dépens

	France	was	ordered	to	bear	its	own	costs
La								
République								
française								
supportera								
ses								
propres								
dépens								

France was ordered to bear its, La République française supportera ses
was ordered to bear its own, supportera ses propres
its own costs, ses propres dépens

	France	was	ordered	to	bear	its	own	costs
La								
République								
française								
supportera								
ses								
propres								
dépens								

France was ordered to bear its own, La République française supportera ses propres

was ordered to bear its own costs, supportera ses propres dépens

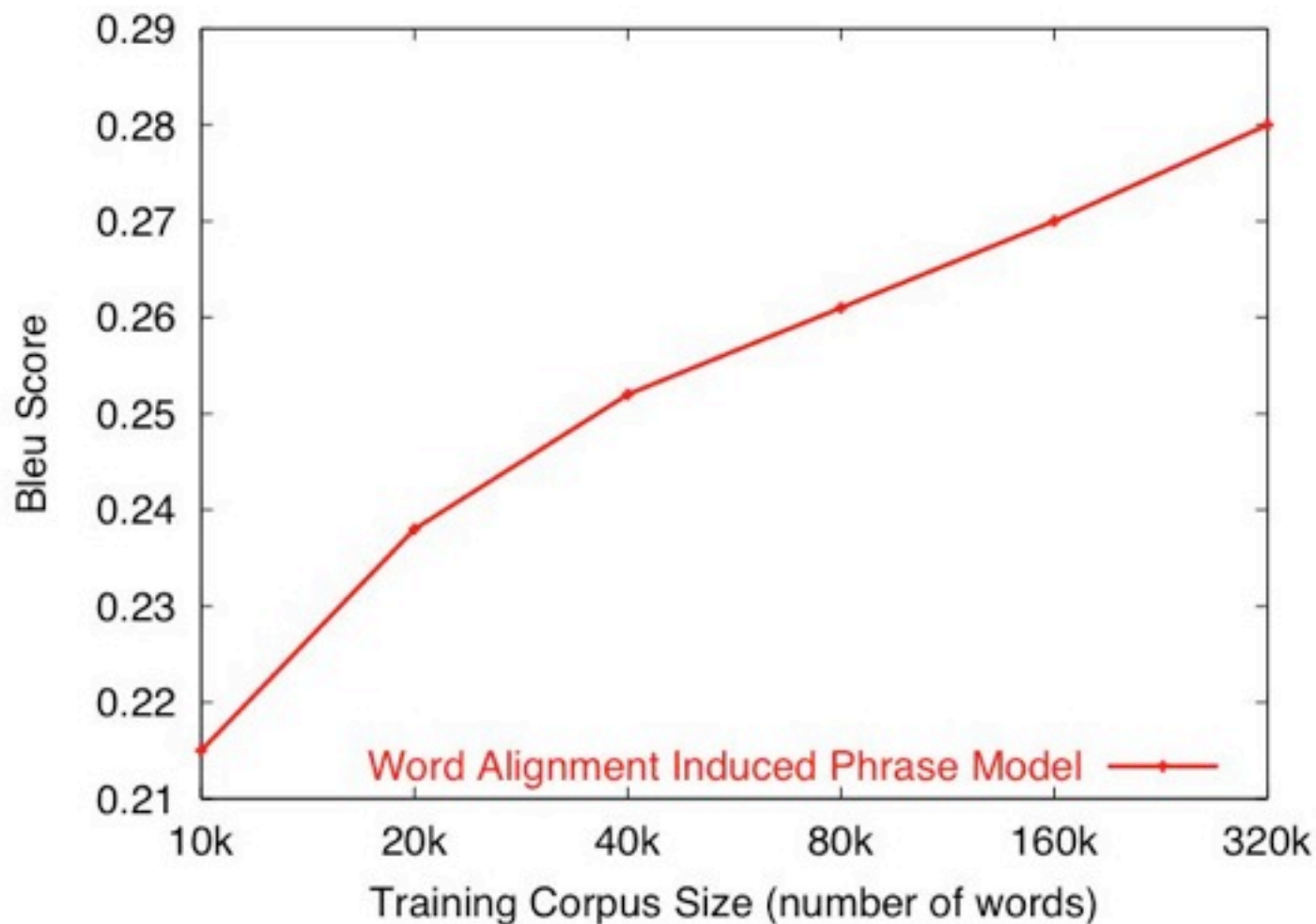
	France	was	ordered	to	bear	its	own	costs
La	■							
République	■							
française	■							
supportera		■	■	■	■			
ses						■		
propres							■	
dépens								■

France was ordered to bear its own costs, La République française supportera ses propres dépens

How we translate

- Break input strings into all possible subphrases
- Look up translations
- Search for most probable recombination

Quality improves with larger corpora



Advantages of SMT (recap)

- Can be applied to any language pair
- Quick to develop
- Improves as more data becomes available
- High-quality phrase-based translation

Linear B's goal: Integration into HT

- Translation memories = parallel corpora
- What if there's not enough data in the TM?
 - *Make system adaptable*
 - *Improve with use through post-editing*

1. 電子メール
/Fax/郵送にて校正
原稿をお送りくださ
い。
2. お見積、納期をお
客様にお知らせいた
し

Linear B Machine Translation

1. Please send a
proofreading
manuscript by the
E-mail / Fax /
mailing.
2. I announce a
visitor an estimate
and time for
delivery.

they post-edit

1. Please send a
manuscript for
proofreading by e-
mail, fax, or post.
2. You will receive
a notification of
receipt containing a
time estimate

Manuscript sent
to client

1. 電子メール
/Fax/郵送にて校正
原稿をお送りくださ
い。
2. お見積、納期をお
客様にお知らせいた
し

1. Please send a
manuscript for
proofreading by e-
mail, fax, or post.
2. You will receive
a notification of
receipt containing a
time estimate

Parallel text sent to
Linear B for retraining

Training time problem

- Training our French-English system with 1.5 million sentences takes 2 weeks
- Solution: *Dynamic updating of our models*

Example

- *Durant l'ère glacière, les changements climatiques obligèrent le règne animale à migrer vers des contrées plus accueillantes.*
- During the era refrigerator, the climatic changes oblige the reign animal to migrate to more accessible regions.
- During the ice age, climactic changes caused the animal kingdom to migrate to more accessible regions.

Weighting contribution

$$p(\bar{f}|\bar{e}) = \frac{\lambda_1 \text{count}_{C_1}(\bar{f}, \bar{e}) + \lambda_2 \text{count}_{C_2}(\bar{f}, \bar{e})}{\lambda_1 \sum_{\bar{f}} \text{count}_{C_1}(\bar{f}, \bar{e}) + \lambda_2 \sum_{\bar{f}} \text{count}_{C_2}(\bar{f}, \bar{e})}$$

- Ensures consistency with user's editing
- May help to port to new domain

Advanced Editing

- Alternative to adding phrases
- Users audit cause of mistranslation
- Edit training data

Example

- *La peine capitale ètè abolie en France sous François Mitterand.*
- The crime punishable by was abolished in France under François Mitterand.

Inspect phrases

La	peine capitale	ètè abolie	en France sous	François Mitterand.
----	----------------	------------	----------------	---------------------

The	crime punishable by	was abolished	in France under	François Mitterand.
-----	---------------------	---------------	-----------------	---------------------

Query and edit

Annotation Tool

In Canada murder is a crime punishable by death

Au	■								
Canada		■							
le			■						
meurtre			■						
est				■					
passible									
de									
peine					■				
capitale						■	■		
.									

Go to 1 of 1

<-- Back Next -->

Annotation Tool

In Canada murder is a crime punishable by death .

Au	■								
Canada		■							
le			■						
meurtre			■						
est				■					
passible									
de									
peine					■	■	■	■	
capitale					■	■	■	■	
.									

Go to 1 of 1

<-- Back Next -->

Conclusion

- Designed editing tools to allow for adaptability and customization
- Makes SMT transparent
- Hope to make SMT more useful in HT

