

Context-Dependent Alignment Models for Statistical Machine Translation

Jamie Brunning, Adrià de Gispert and William Byrne

Department of Engineering
Cambridge University

NAACL HLT, Boulder, Colorado, 1st June 2009

Overview

- Introduction
 - Motivate use of context
 - Introduce context-dependent alignment models
 - Using part-of-speech as context
- Context-dependent alignment models
 - Decision tree clustering of contexts
 - Questions for decision trees
 - Model training
- Experimental results
 - Evaluation of alignment quality
 - Translation results

Introduction

- Current alignment models (e.g. Model 1, Model 2, HMMs) rely on context-independent word-to-word translation probabilities
 - Translation probability distribution does not depend on context
- We introduce translation dependency based on the context of the source word
 - Neighbouring words should influence translation
- Describe modelling methods to capture context
- Introduce clustering to control model complexity
- Evaluate models using Alignment Error Rate (AER) and BLEU score of translation for Arabic-English and Chinese-English

Context-dependent translation

- Given data:

Arabic context sequence	$d_1 d_2 \dots d_J$
Arabic sentence	$f_1 f_2 \dots f_J$
English sentence	$e_1 e_2 \dots e_I$
English context sequence	$c_1 c_2 \dots c_I$
- Models are extended to use context-dependent word-to-word translation probabilities, i.e. $P(f_1^J | e_1^I)$ is replaced by $P(f_1^J | e_1^I, c_1^I)$
- For Model 1,

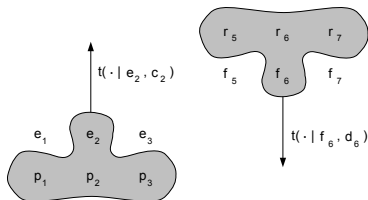
$$\text{Context-independent Model 1} \quad \frac{1}{(J+1)^I} \prod_{i=1}^I t(e_i | f_{a_i})$$

$$\text{Context-dependent Model 1} \quad \frac{1}{(J+1)^I} \prod_{i=1}^I t(e_i | f_{a_i}, c_{a_i}).$$

- The same can be done for Model 2 and HMM
- Alignment components of models remain the same
- The models can still be trained using EM

Using part-of-speech tags as context

- We have POS tags for both languages:
 - Arabic tags $r_1 r_2 \dots r_J$
 - Arabic $f_1 f_2 \dots f_J$
 - English $e_1 e_2 \dots e_I$
 - English tags $p_1 p_2 \dots p_I$
- Define context to be the part-of-speech tags of previous, current and next source words: $c_i = p_{i-1}, p_i, p_{i+1}$,



$$d_j = r_{j-1}, r_j, r_{j+1}$$

- Assume each word has single context which is given - could be a probability distribution over contexts but this is not considered

Decision tree clustering

- Without clustering, computational complexity of models is too large and models suffer from data sparsity
- One decision tree per word in the source vocabulary
 - Assume that different contexts affect each word in a different way
- For each source word, begin with root node containing all contexts of the word
 - 1 Compute the value of the quality function at each leaf node
 - 2 For each leaf node, find the question \hat{q} which produces the largest increase in quality function when used to split that node
 - Subject to constraints on improvement and occupancy
 - 3 Split the node with largest increase in quality function
 - 4 Repeat steps 1 – 3 until no leaf node satisfies the constraints

Quality function for clustering

- During clustering, we locally maximise EM auxiliary function
- For a source word e and cluster K of contexts of e , this leads to quality function

$$Q_e(K) = \sum_f \sum_{c \in K} \gamma'(f|e, c) \log t(f|e, K)$$

- $\gamma'(f|e, c)$ is the EM count of e with context c translating to f
- γ' initially computed from the CI model
- All contexts in a cluster have the same translation probability distribution
- Distribution for cluster K is given by

$$t(f|e, K) = \frac{\sum_{c \in K} \gamma'(f|e, c)}{\sum_{f'} \sum_{c \in K} \gamma'(f'|e, c)}$$

Quality function for clustering II

$$Q_e(K) = \sum_f \sum_{c \in K} \gamma'(f|e, c) \log t(f|e, K)$$

- Look at the effect of splitting cluster K using question q ; we choose \hat{q} to maximise the increase in quality

$$\hat{q} = \arg \max_q Q_e(K_q) + Q_e(K_{\bar{q}}) - Q_e(K)$$

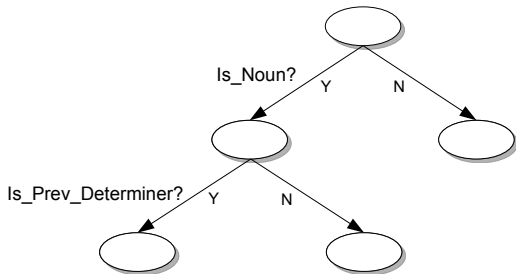
- $K_q = \{\text{contexts that answer 'yes' to } q\}$,
 $K_{\bar{q}} = \{\text{contexts that answer 'no' to } q\}$

Choosing questions

- Framework allows for general questions of the form *'Is the previous word x ?' and 'Does word y occur within n words of this one?'*
 - Having too many distinct questions leads to computational intractability
- Restrict the questions to those concerning the part-of-speech tag assigned to the current, previous and next source words.
- For each part-of-speech tag T , we ask *'Does w have tag T ?'.*
- Group part-of-speech tags to ask more general questions: e.g. *'Is w a noun?'*
- Ask the same questions of the previous and next words in the source sentence
- Total of 152 distinct questions in English; 68 questions in Arabic; 100 questions in Chinese.

Example decision tree

- Decision tree for English word *endeavour*
endeavour



he assured that he would endeavour to provide an answer .
the participation of the people concerned is an integral
part in this endeavour .
they assign high priority to their cooperation in these
areas of endeavour .

Model training

- In each alignment direction:
 - Context-independent models are trained for several EM iterations
 - Clustering of source word contexts carried out using statistics from CI models
 - Context-dependent models with clustered source contexts initialised from CI models
 - Clustering and EM algorithm parallelized for efficiency
 - More iterations of EM training performed on the context-dependent models
- Models are trained in both alignment directions
- Alignments from each direction are merged in the usual way prior to extraction of rules for a hierarchical phrase-based translation system

Data sets and text processing

- Arabic-English
 - NIST MT08 Arabic-English parallel text excluding UN collections
 - MADA toolkit used to perform Arabic morphological word decomposition and POS tagging
- Chinese-English
 - LDC Chinese-English parallel text allowed for NIST MT08
 - Chinese text tagged using MXPOST maximum entropy tagger trained on the Penn Chinese Treebank 5.1
- All English text tagged using TnT tagger trained on WSJ portion of Penn treebank

Data set	Sentences	Foreign words	English words
Arabic-English	300k	8.4M	9.5M
Chinese-English	8.2M	207M	220M

Frequently used root node questions

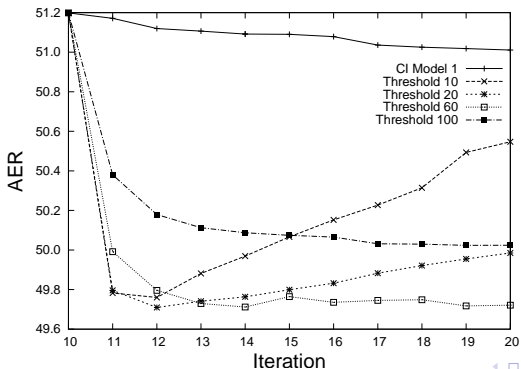
- Questions that individually have the greatest ability to discriminate between clusters that translate differently
 - Often depends on previous word: of 50 most common in English, 25 questions are about the previous word, 19 about the next, 6 about the current word

English question	Frequency
Is_Next_Preposition	1523
Is_Prev_Determiner	1444
Is_Prev_Preposition	1209
Is_Prev_Adjective	864
Is_Next_Noun_Singular_Mass	772

Arabic question	Frequency
Is_Prev_Preposition	1110
Is_Next_Preposition	993
Is_Prev_Noun	981
Is_Next_Noun	912
Is_Prev_Coordinating_Conjunction	627

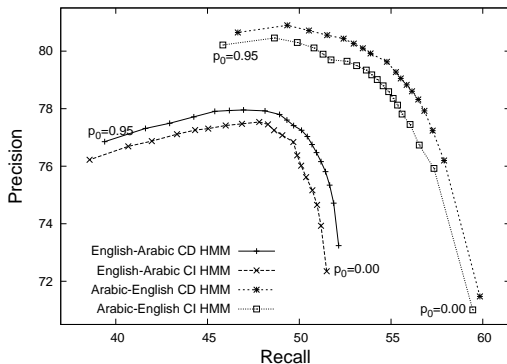
Model 1 alignment quality

- Context-dependence improves AER of Model 1 in English-Arabic direction
 - Similar patterns for Arabic-English
 - Significant drop in AER over first 2 iterations of training
 - Some evidence of overtraining for lower thresholds



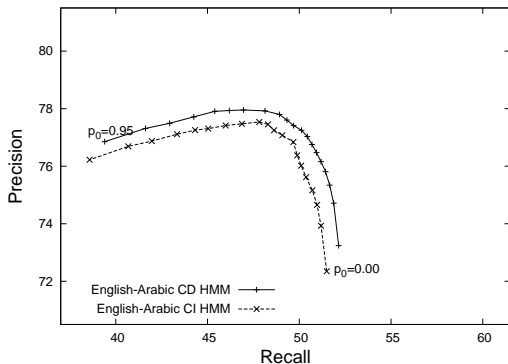
HMM Alignment quality

- Context-dependent HMMs give improved alignment quality over CI HMMs, demonstrated by precision-recall curve
 - For given precision, recall of CD HMM is larger; for given recall, precision of CD HMM is larger



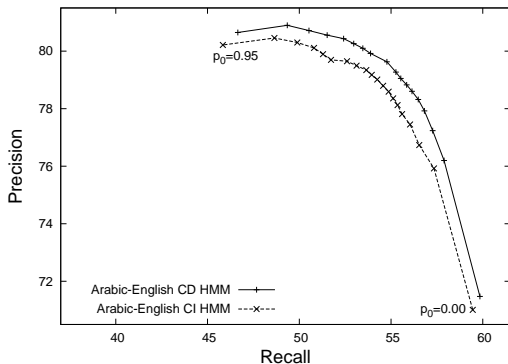
HMM Alignment quality

- Context-dependent HMMs give improved alignment quality over CI HMMs, demonstrated by precision-recall curve
 - For given precision, recall of CD HMM is larger; for given recall, precision of CD HMM is larger



HMM Alignment quality

- Context-dependent HMMs give improved alignment quality over CI HMMs, demonstrated by precision-recall curve
 - For given precision, recall of CD HMM is larger; for given recall, precision of CD HMM is larger



Translation results

- State-of-the-art hierarchical decoder used - HiFST (Iglesias et al. 2009)
- Tuning carried out using MET on held-out data set
- Evaluation by BLEU score

Arabic-English			
Alignments	tune	mt02_05_test	MT08-nw
CI HMM	50.0	49.4	46.3
CD HMM	50.0	49.7	46.9

Chinese-English full NIST MT08 parallel text			
Alignments	tune	test-nw	MT08-nw
CI HMM	30.8	31.2	29.0
CD HMM	31.1	31.5	29.8

- Decoding carried out using 4-gram language model in first pass
 - Gains are still seen when lattices of translation hypotheses are rescored using large 5-gram language models

Conclusion

- Introduced context-dependent Model 1 and HMM alignment models
 - Context information is incorporated by the use of part-of-speech tags in both languages of the parallel text
 - Developed decision tree clustering algorithms to cluster source word contexts based on optimisation of the EM auxiliary function
- Context-dependent alignment models improve alignment quality, and the quality of machine translation systems built on those alignments
- Will be made available in a new release of MTTK
 - <http://mi.eng.cam.ac.uk/~wjb31/distrib/mttkv1/>

Thank you!

- Any questions?

<http://mi.eng.cam.ac.uk/~wjb31/distrib/mttkv1/>

Related work

- Toutanova et al. (2002) use part-of-speech information to estimate alignment probabilities
 - Not used to estimate translation probabilities
- Popović and Ney (2004) use base form of word and POS tag during estimation of word-to-word translation probabilities
 - No context-dependent probabilities using previous or next words
- Stroppa et al. (2007) consider context-informed features as components of the log-linear model during phrase-based translation
 - Not used in alignment models

Future work

- Clustering the probability distributions of infrequently occurring may produce improvements in alignment quality
- Different model training schemes and extensions of the context-dependence to more sophisticated alignment models
- Further translation experiments

Bibliography



M. Popović and H. Ney.

Improving word alignment quality using morpho-syntactic information.

In *In Proceedings of COLING*, page 310, 2004.

doi: <http://dx.doi.org/10.3115/1220355.1220400>.



N. Stroppa, A. van den Bosch, and A. Way.

Exploiting source similarity for SMT using context-informed features.

In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, pages 231 – 240, 2007.



K. Toutanova, H. T. Ilhan, and C. D. Manning.

Extensions to HMM-based statistical word alignment models.

In *Proceedings of EMNLP*, pages 87–94, 2002.