Improved Word Alignments for Statistical Machine Translation
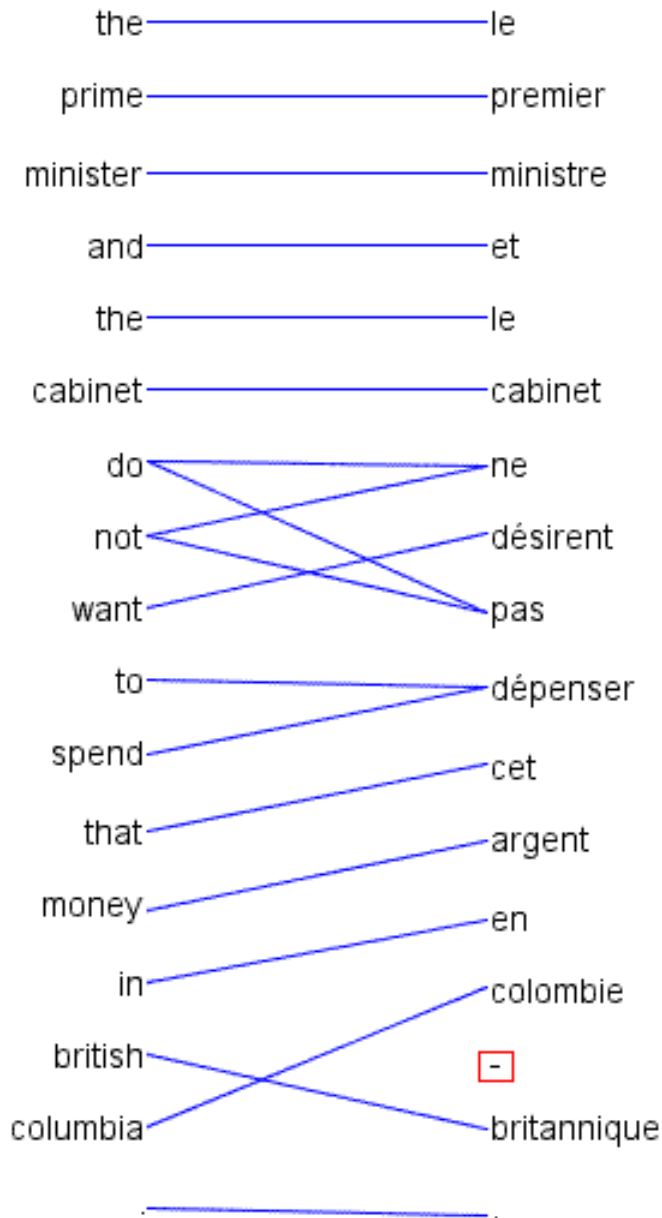
Alex Fraser

Institute for NLP

University of Stuttgart

Statistical Machine Translation (SMT)

- Build a model P(e | f), the probability of the English sentence "e" given the French sentence "f"
- To translate a French sentence "f", choose the English sentence "e" which maximizes P(e | f)

$$\text{argmax}\ P(\ e\ |\ f\ )\ =\ \text{argmax}\ P(\ f\ |\ e\ )\ P(\ e\ )$$
$$\quad\ e \qquad\qquad\qquad\qquad e$$

- P(f | e) is the "translation model"
  - Collect statistics from word aligned parallel corpora
- P(e) is the "language model"

Alex Fraser

Annotation of Minimal Translational Correspondences

the ——————— le
prime ——————— premier
minister ——————— ministre
and ——————— et
the ——————— le
cabinet ——————— cabinet
do ——————— ne
not ——————— désirent
want ——————— pas
to ——————— dépenser
spend ——————— cet
that ——————— argent
money ——————— en
in ——————— colombie
british ——————— -
columbia ——————— britannique
. ——————— .

- Word alignment is annotation of minimal translational correspondences

- Annotated in the context in which they occur

- Not idealized translations!

(solid blue lines annotated by a bilingual expert)

Alex Fraser

Overview

- Solving problems with previous word alignment methodologies
  - Problem 1: Measuring quality
  - Problem 2: Modeling
  - Problem 3: Utilizing new knowledge
  - Joint Work with Daniel Marcu, USC/ISI

Alex Fraser

Problem 1: Existing Metrics Do Not Track Translation Quality

- Dozens of papers report word alignment quality increases according to intrinsic metrics

- Contradiction: few of these report MT results; those that do report inconclusive gains

- This is because the two commonly used intrinsic metrics, AER and balanced F-Measure, do not correlate with MT performance!
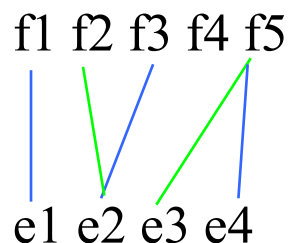
Measuring Precision and Recall

- Start by fully linking hypothesized alignments
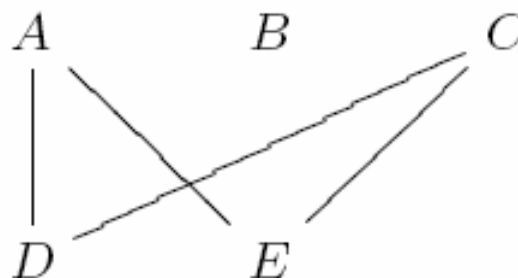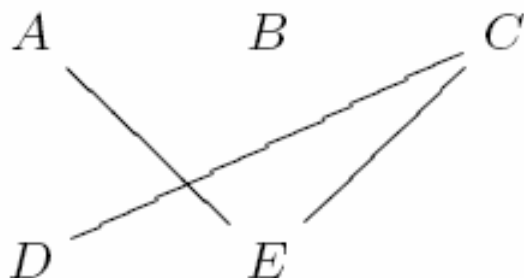
- Precision is the number of links in our hypothesis that are correct
  - If we hypothesize there are no links, have 100% precision
- Recall is the number of correct links we hypothesized
  - If we hypothesize all possible links, have 100% recall
- We will test metrics which formally define and combine these in different ways
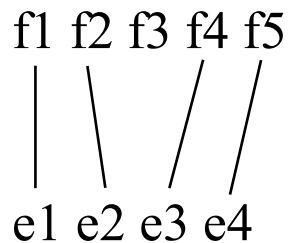
Alignment Error Rate (AER)

Gold

f1 f2 f3 f4 f5

e1 e2 e3 e4

Hypothesis

f1 f2 f3 f4 f5

e1 e2 e3 e4

$$Precision(A, P) = \frac{|P \cap A|}{|A|} = \frac{3}{4}$$

(e3,f4) wrong

$$Recall(A, S) = \frac{|S \cap A|}{|S|} = \frac{2}{3}$$

(e2,f3) not in hyp

$$AER(A, P, S) = 1 - \frac{|P \cap A| + |S \cap A|}{|S| + |A|} = \frac{2}{7}$$

BLUE = sure links   GREEN = possible links

Alex Fraser

Experiment

- Desideratum:
  - Keep everything constant in a set of SMT systems except the word-level alignments
    - Alignments should be realistic

- Experiment:
  - Take a parallel corpus of 8M words of Foreign-English. Word-align it. Build SMT system. Report AER and Bleu.
  - For better alignments: train on 16M, 32M, 64M words (but use only the 8M words for MT building).
  - For worse alignments: train on $2 \times 1/2$, $4 \times 1/4$, $8 \times 1/8$ of the 8M word training corpus.

- If AER is a good indicator of MT performance, $1 - $ AER and BLEU should correlate no matter how the alignments are built (union, intersection, refined)
  - Low $1 - $ AER scores should correspond to low BLEU scores
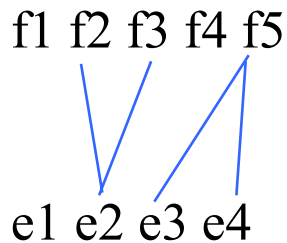 - High $1 - $ AER scores should correspond to high BLEU scores

AER is not a good indicator of MT performance

$r^2 = 0.16$

F$_\alpha$-score

Gold

f1 f2 f3 f4 f5

e1 e2 e3 e4

Hypothesis

f1 f2 f3 f4 f5

e1 e2 e3 e4
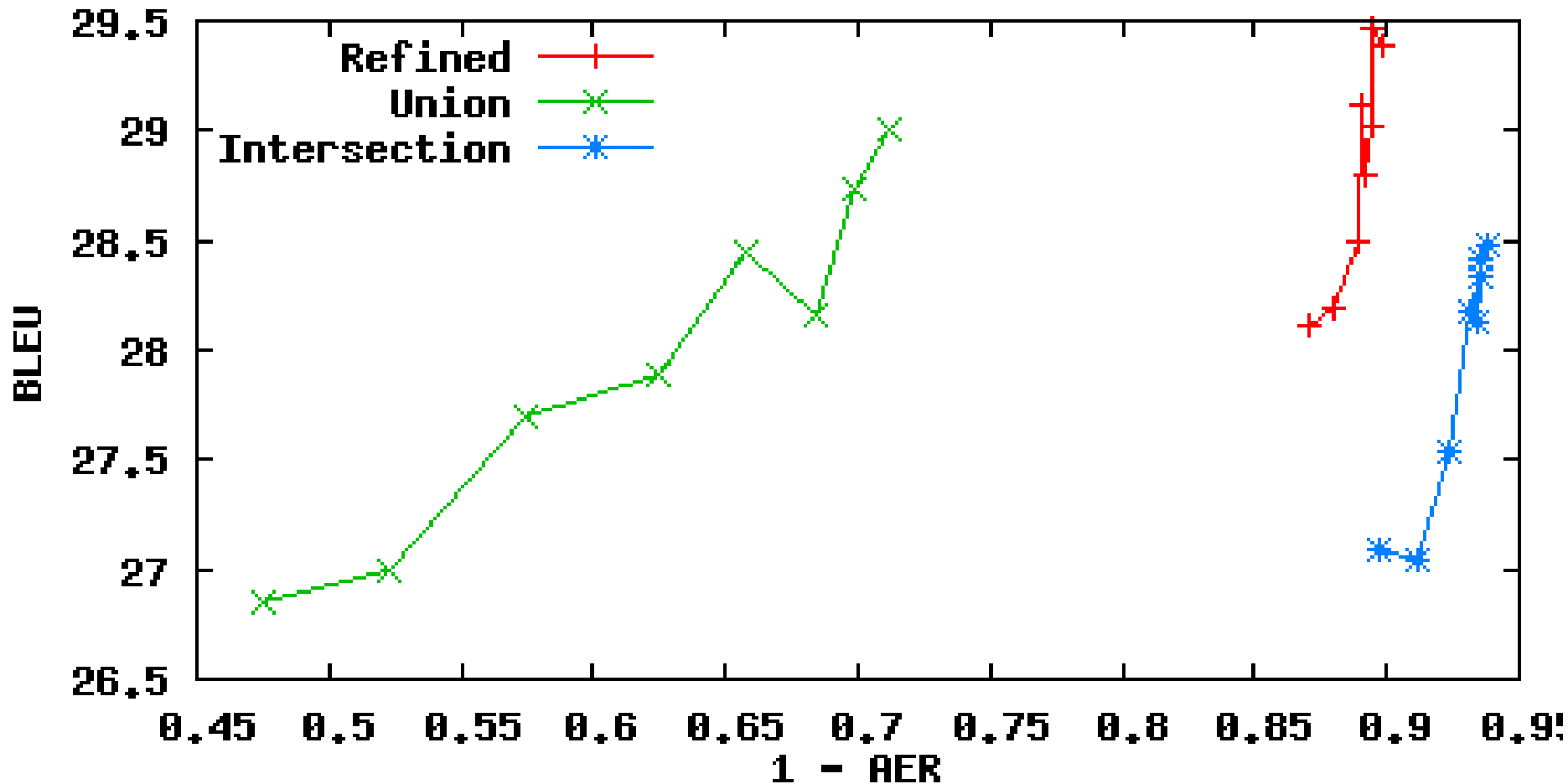
$$Precision(A, S) = \frac{|S \cap A|}{|A|} = \frac{3}{4}$$

(e3,f4)
wrong

$$Recall(A, S) = \frac{|S \cap A|}{|S|} = \frac{3}{5}$$

(e2,f3)
(e3,f5)
not in hyp
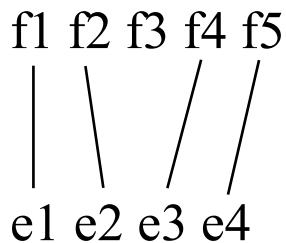
$$F(A, S, \alpha) = \frac{1}{\dfrac{\alpha}{Precision(A,S)} + \dfrac{1-\alpha}{Recall(A,S)}}$$

Called F$_\alpha$-score to differentiate
from ambiguous term F-Measure

Alex Fraser

F$_\alpha$-score is a good indicator of MT performance

$r^2 = 0.85$ $\alpha = 0.4$

Alex Fraser
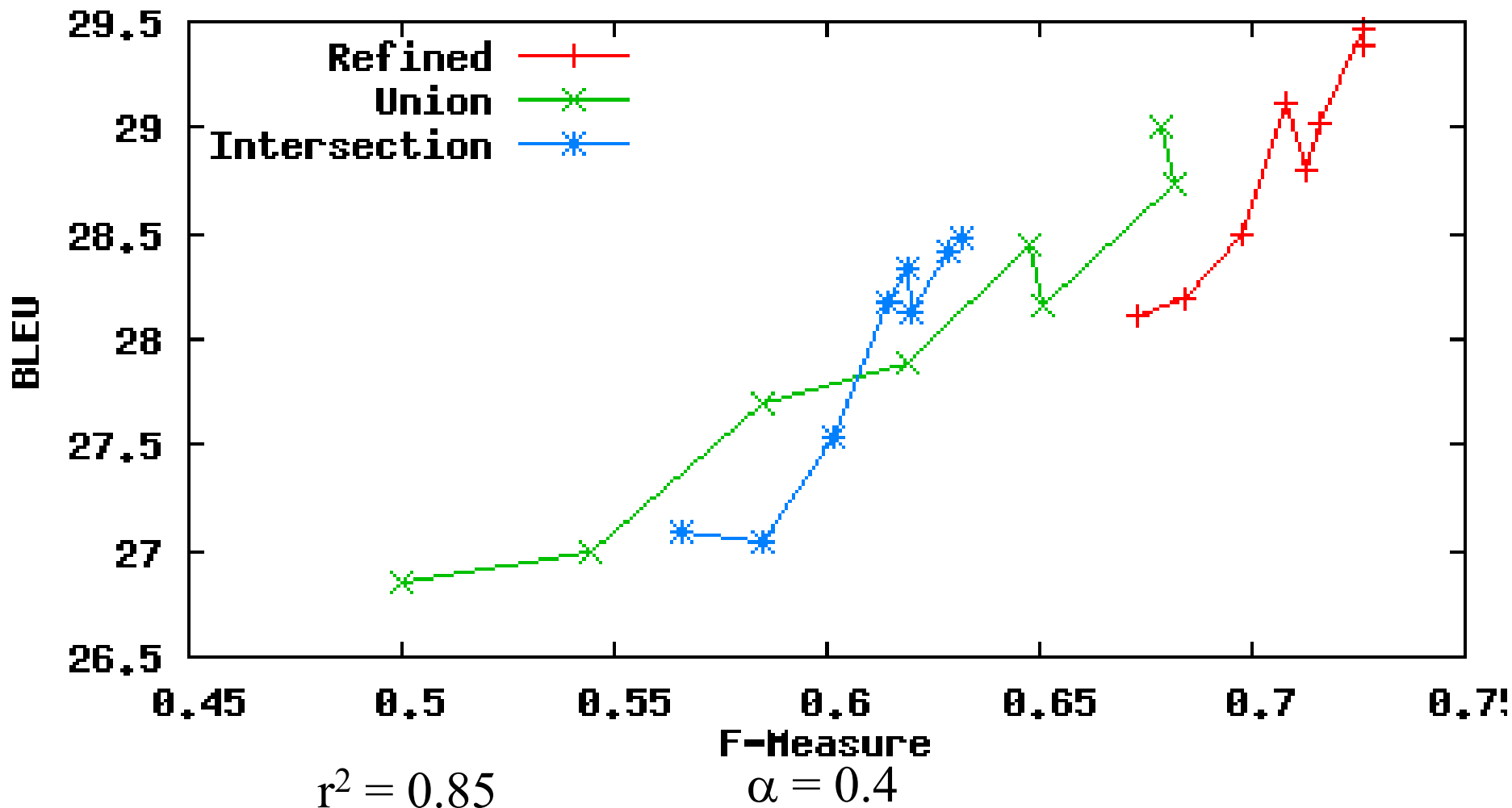
Discussion

- Using F_α-score as a loss criterion will allow for development of discriminative models (later in talk)

- AER is not derived correctly from F-Measure

- For details of experiments see squib in Sept. 2007 Computational Linguistics

Problem 2: Modeling the Wrong Structure
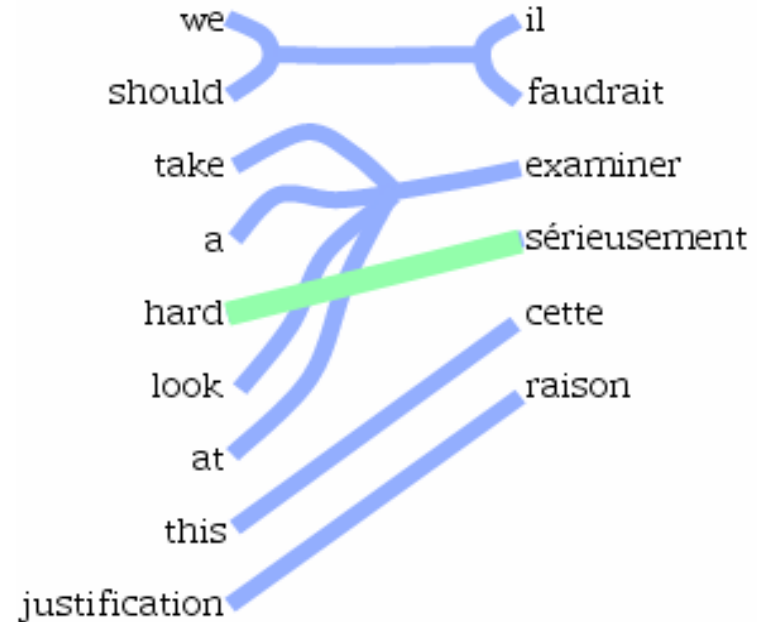
- 1-to-N assumption
  - Multi-word "cepts" (words in one language translated as a unit) only allowed on target side. Source side limited to single word "cepts".
- Phrase-based assumption
  - "cepts" must be consecutive words

LEAF Generative Story

	source	absolutely	[comma]	they	do	not	want	to	spend	that	money
word type (1)		DEL.	DEL.	HEAD	non-head	HEAD	HEAD	non-head	HEAD	HEAD	HEAD
linked from (2)				THEY	do	NOT	WANT	to	SPEND	THAT	MONEY
head(3)				ILS		PAS	DESIRENT		DEPENSER	CET	ARGENT
cept size(4)				1		2	1		1	1	1
num spurious(5)		1									
spurious(6)		aujourd'hui									
non-head(7)				ILS	PAS	ne	DESIRENT		DEPENSER	CET	ARGENT
placement(8)		aujourd'hui		ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT
spur. placement(9)				ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT	aujourd'hui

- Explicitly model three word types:
  - **Head word**: provide most of conditioning for translation
    - Robust representation of multi-word cepts (for this task)
    - This is to semantics as ``syntactic head word'' is to syntax
  - **Non-head word**: attached to a head word
  - **Deleted source words** and **spurious target words** (NULL aligned)

Alex Fraser

LEAF Generative Story

source	absolutely	[comma]	they	do	not	want	to	spend	that	money
word type (1)	DEL.	DEL.	HEAD	non-head	HEAD	HEAD	non-head	HEAD	HEAD	HEAD
linked from (2)			THEY	do	NOT	WANT	to	SPEND	THAT	MONEY
head(3)			ILS		PAS	DESIRENT		DEPENSER	CET	ARGENT
cept size(4)			1		2	1		1	1	1
num spurious(5)	1									
spurious(6)	aujourd'hui									
non-head(7)			ILS	PAS	ne	DESIRENT		DEPENSER	CET	ARGENT
placement(8)	aujourd'hui		ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT
spur. placement(9)			ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT	aujourd'hui

- Once source cepts are determined, exactly one target head word is generated from each source head word

- Subsequent generation steps are then conditioned on a single target and/or source head word
- See EMNLP 2007 paper for details

LEAF

- Can score the same structure in both directions
- Math in one direction (please do not try to read):

$$p(f, a|e) = [\prod_{i=1}^{l} g(\chi_i|e_i)]$$

$$[\prod_{i=1}^{l} \delta(\chi_i, -1)w_{-1}(\mu_i - i|\mathsf{class}_e(e_i))]$$

$$[\prod_{i=1}^{l} \delta(\chi_i, 1)t_1(\tau_{i1}|e_i)][\prod_{i=1}^{l} \delta(\chi_i, 1)s(\psi_i|e_i, \gamma_i)]$$

$$[s_0(\psi_0|\sum_{i=1}^{l} \psi_i)][\prod_{k=1}^{\psi_0} t_0(\tau_{0k})]$$

$$[\prod_{i=1}^{l} \prod_{k=2}^{\psi_i} t_{>1}(\tau_{ik}|e_i, \mathsf{class}_h(\tau_{i1}))]$$

$$[\prod_{i=1}^{l} \prod_{k=1}^{\psi_i} D_{ik}(\pi_{ik})]$$

Discussion

- LEAF is a powerful model
- But, exact inference is intractable
  - We use hillclimbing search from an initial alignment
- First model of correct structure: M-to-N discontiguous
  - Head word assumption allows use of multi-word cepts
    - Decisions robustly decompose over words
    - Does not have segmentation problem of phrase alignment models: Probability of alignments of cept "the man" are closely related to probabilities for cept "man"
  - Not limited to only using 1-best prediction

Problem 3: Existing Approaches Can't Utilize New Knowledge

- It is difficult to add new knowledge sources to generative models
    - Requires completely reengineering the generative story for each new source

- Existing unsupervised alignment techniques can not use manually annotated data

Background

- We love EM, but
  - EM often takes us to places we never imagined/wanted to go
- Bayes is always right

$$\underset{e}{\text{argmax}} \ \ P(e\,|\,f) \ = \ \underset{e}{\text{argmax}} \ \ P(e) \text{ x } P(f\,|\,e)$$

But in practice, this works better:

$$\underset{e}{\text{argmax}} \ \ P(e)^{2.4} \text{ x } \ P(f\,|\,e) \text{ x length}(e)^{1.1} \text{ x KS }^{3.7} \dots$$

Decomposing LEAF

- Decompose each step of the LEAF generative story into a sub-model of a log-linear model
  - Add backed off forms of LEAF sub-models
  - Add heuristic sub-models (do not need to be related to generative story!)
  - Allows tuning of vector λ which has a scalar for each sub-model controlling its contribution

Alex Fraser

Reinterpreting LEAF

- $g(e_i)$ – source word type sub-model
- $w(\mu_i)$ – source non-head linking sub-model
- $t_1(f_j \mid y(i))$ – head word translation sub-model

- Etc…   – many more sub-models

$$p(a, f \mid e) = g \times w \times t_1 \times \text{etc…}$$

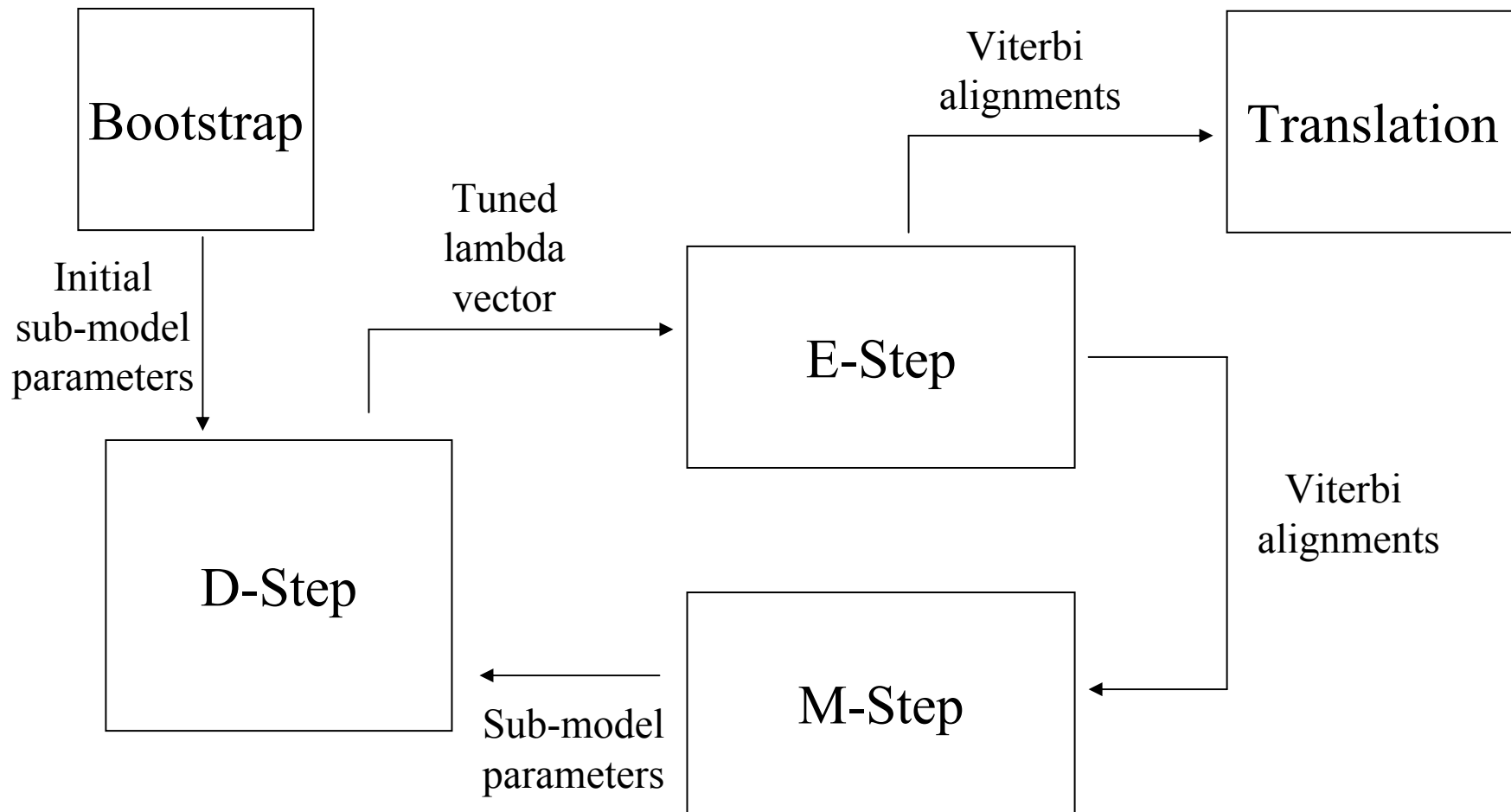$$p(a, f \mid e) = z^{-1} \times g^{\lambda 1} \times w^{\lambda 2} \times t_1{}^{\lambda 3} \times \text{etc…}$$

$$p(a, f \mid e) = \frac{\exp \sum_m \lambda_m h_m(f, a, e; \theta_m)}{\exp(Z)}$$

Alex Fraser

Semi-Supervised Training

- Define a semi-supervised algorithm which alternates increasing likelihood with decreasing error
  - Increasing likelihood is similar to EM
  - Discriminatively bias EM to converge to a local maxima of likelihood which corresponds to "better" alignments
    - "Better" = higher F_α-score on small gold standard corpus

The EMD Algorithm

Bootstrap

Initial sub-model parameters

Tuned lambda vector

E-Step

Viterbi alignments

Translation

Viterbi alignments

M-Step

Sub-model parameters

D-Step

Alex Fraser

Discussion

- Usual formulation of semi-supervised learning: "using unlabeled data to help supervised learning"
  - Build initial supervised system using labeled data, predict on unlabeled data, then iterate
  - But we do not have enough gold standard word alignments to estimate parameters directly!
- EMD allows us to train a small number of important parameters discriminatively, the rest using likelihood maximization, and allows interaction
  - Similar in spirit (but not details) to semi-supervised clustering

Experiments

- French/English
  - LDC Hansard (67 M English words)
  - MT: Alignment Templates, phrase-based
- Arabic/English
  - NIST 2006 task (168 M English words)
  - MT: Hiero, hierarchical phrases

Results

System	French/English		Arabic/English	
	F-Measure (α = 0.4)	BLEU (1 ref)	F-Measure (α = 0.1)	BLEU (4 refs)
IBM Model 4 (GIZA++) and heuristics	73.5	30.63	75.8	51.55
EMD (ACL 2006 model) and heuristics	74.1	31.40	79.1	52.89
LEAF+EMD	76.3	31.86	84.5	54.34

Alex Fraser

Contributions

- Found a metric for measuring alignment quality which correlates with MT quality

- Designed LEAF, the first generative model of M-to-N discontiguous alignments

- Developed a semi-supervised training algorithm, the EMD algorithm

- Obtained large gains of 1.2 BLEU and 2.8 BLEU points for French/English and Arabic/English tasks

Thank You!