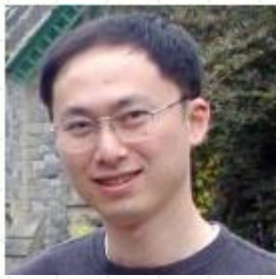
SPEECH TRANSLATION: THEORY AND PRACTICES

Bowen Zhou and Xiaodong He



zhou@us.ibm.com
xiaohe@microsoft.com

May 27th 2013 @ ICASSP 2013

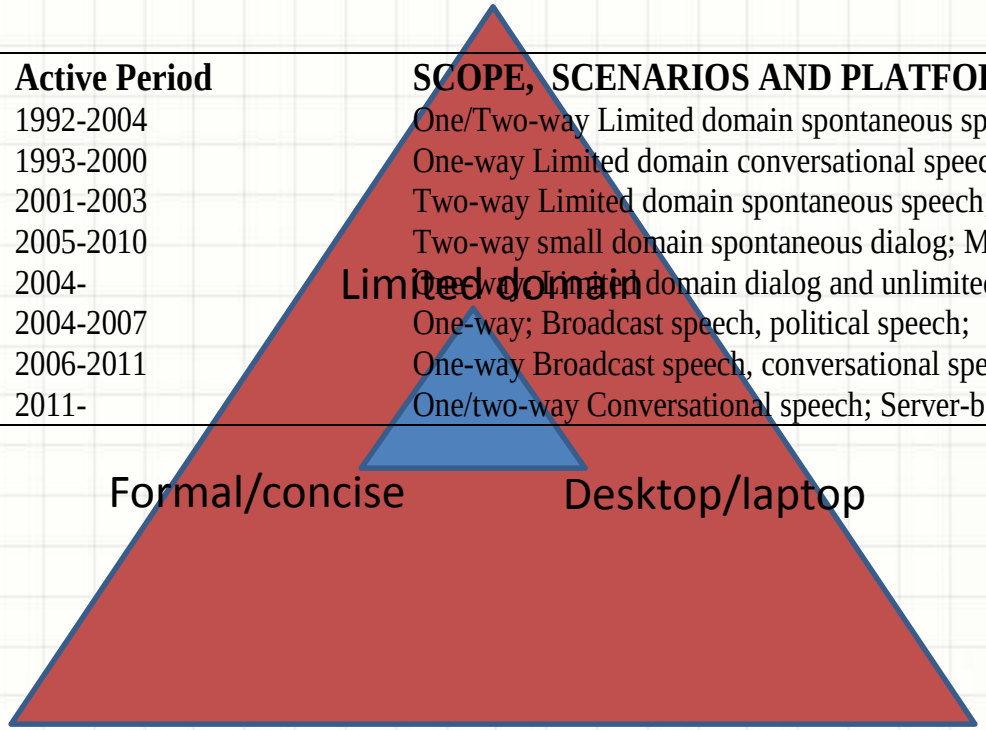
Universal Translator: dream (will) come true *or* yet another over promise ?

- Spoken communication without a language barrier: mankind's long-standing dreams
  - Translate human speech in one language into another (text or speech)
  - Automatic Speech Recognition: ASR
  - (Statistical) Machine Translation: (S)MT
- We've been promised by Sci-fi (*Star Trek)* for 5 decades
- Serious research efforts have started in early 1990s'
  - When statistical ASR (e.g., HMM-based) starts showing dominance and SMT was emerging

ST Research Evolves Over 2 Decades

Broad domains/large vocabulary

PROJECT/CAMPAIGN	Active Period	SCOPE, SCENARIOS AND PLATFORMS
C-STAR	1992-2004	One/Two-way Limited domain spontaneous speech
VERBMOBIL	1993-2000	One-way Limited domain conversational speech
DARPA BABYLON	2001-2003	Two-way Limited domain spontaneous speech; Laptop/Handheld
DARPA TRANSTAC	2005-2010	Two-way small domain spontaneous dialog; Mobile/Smartphone;
IWSLT	2004-	One-way Limited domain dialog and unlimited free-style talk
TC-STAR	2004-2007	One-way; Broadcast speech, political speech; Server-based
DARPA GALE	2006-2011	One-way Broadcast speech, conversational speech; Server-based
DARPA BOLT	2011-	One/two-way Conversational speech; Server-based

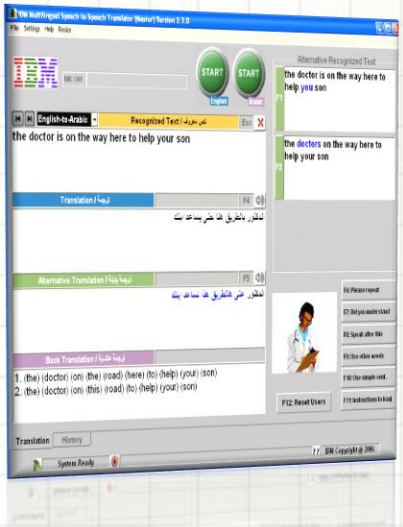


Conversational/Interactive

Mobile/Cloud

Are we closer today?

- Speech translation demos:
  - Completely smartphone-hosted speech-to-speech translation (IBM Research)
  - Rashid's live demo (MSR) in 21st CCC keynote
- Statistical approaches to both ASR/SMT are one of the keys for the success
  - Large data, discriminative training, better models (e.g., DNN for ASR) etc.



Our Objectives

1. Review & analyze state-of-the-art SMT theory, *from ST's perspective*
2. Unifying ASR and SMT to catalyze joint ASR/SMT for improved ST
3. ST's practical issues & future research topics

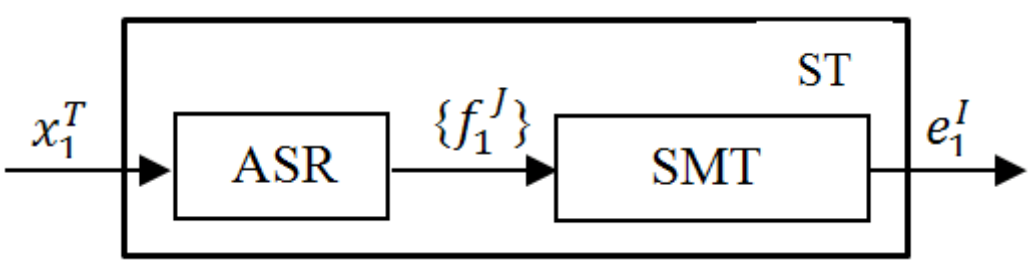
In this talk…

1. Overview: ASR, SMT, ST, Metric
2. Learning Problems in SMT
3. Translation Structures for ST [20 min coffee Break]
4. Decoding: A Unified Perspective ASR/SMT
5. Coupling ASR/SMT: Decoding & Modeling
6. Practices
7. Future Directions

- Technical papers accompanying this tutorial :
  - Bowen Zhou, Statistical Machine Translation for Speech: A Perspective on Structure, Learning and Decoding, in *Proceedings of the IEEE*, IEEE, May 2013
  - Xiaodong He and Li Deng, Speech-Centric Information Processing: An Optimization-Oriented Approach, in *Proceedings of the IEEE*, IEEE, May 2013
- Papers are available on authors' webpages
- Charts coming up soon

Overview & Backgrounds

Speech Translation Process



- Input: a source speech signal sequence $x_1^T = x_1, \cdots, x_T$
- ASR: recognizes it as a set of source word sequences, $\{f_1^J = f_1, \cdots, f_J\}$
- SMT: Translated into the target language sequence of words $e_1^I = e_1, \cdots, e_I$

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} P(e_1^I | x_1^T)$$

$$= \underset{e_1^I}{\operatorname{argmax}} \left\{ \sum_{f_1^J} P(e_1^I | f_1^J) P(x_1^T | f_1^J) P(f_1^J) \right\}$$

First comparison: ASR vs. SMT

- Connection: sequential pattern recognition
  - Determine a sequence of symbols that is regarded as the optimal equivalent in the target domain
    - ASR: $x_1^T \rightarrow f_1^J$
    - SMT $f_1^J \rightarrow e_1^I$.
  - Hence, many techniques are closely related.
- Difference: ASR is monotonic but SMT is not
  - Different modeling/decoding formalisms required
  - One of the key issues we address in this tutorial

ASR in a Nutshell

ASR Decoding

- $\hat{f}_1^J = \underset{f_1^J}{\operatorname{argmax}} P(x_1^T | f_1^J) P(f_1^J)$

Acoustic Models (HMM)

- $P(x_1^T | f_1^J) = \sum_{q_1^T} \prod_t p(q_t | q_{t-1}) \prod_t p(x_t | q_t)$
- $p(x_t | q_t)$ by Gaussian Mixture Models or NN

Language Models (N-gram)

- $p(f_1^J) \approx \prod_{j=1}^J p(f_j | f_{j-N+1}, \cdots, f_{j-1})$

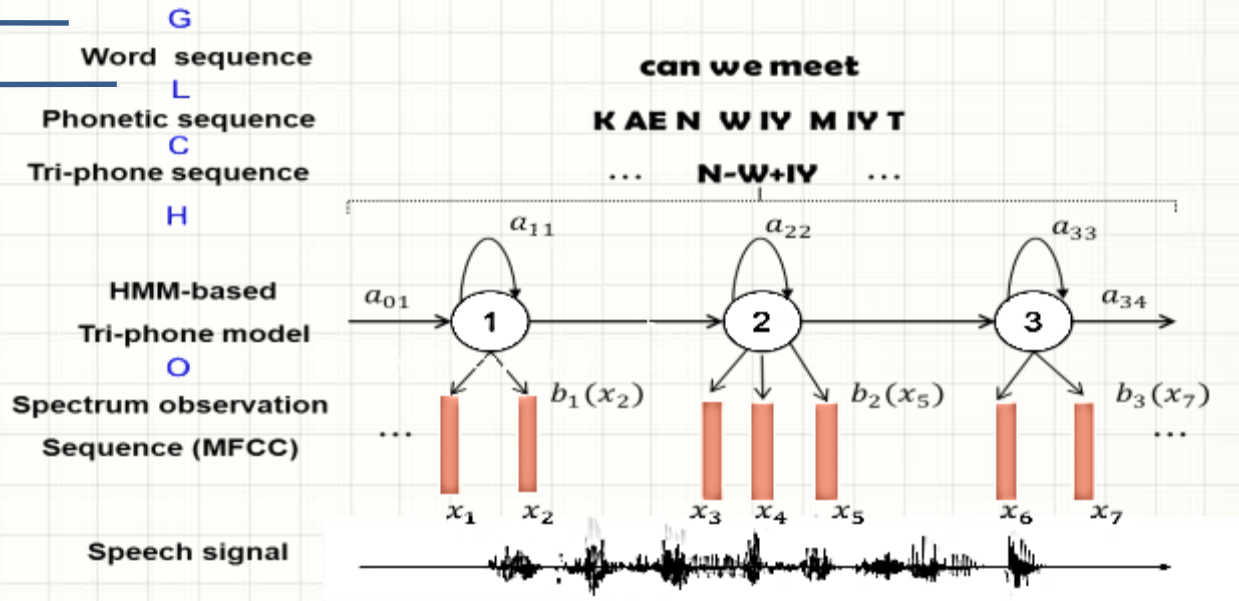
ASR Structures: A Finite-State Problem

Search on a composed finite-state space (graph)

$$\hat{f}_1^J = \text{best\_path}\,(\boldsymbol{O} \circ \boldsymbol{H} \circ \boldsymbol{C} \circ \boldsymbol{L} \circ \boldsymbol{G})$$

G: a weighted acceptor that assigns language model probabilities.

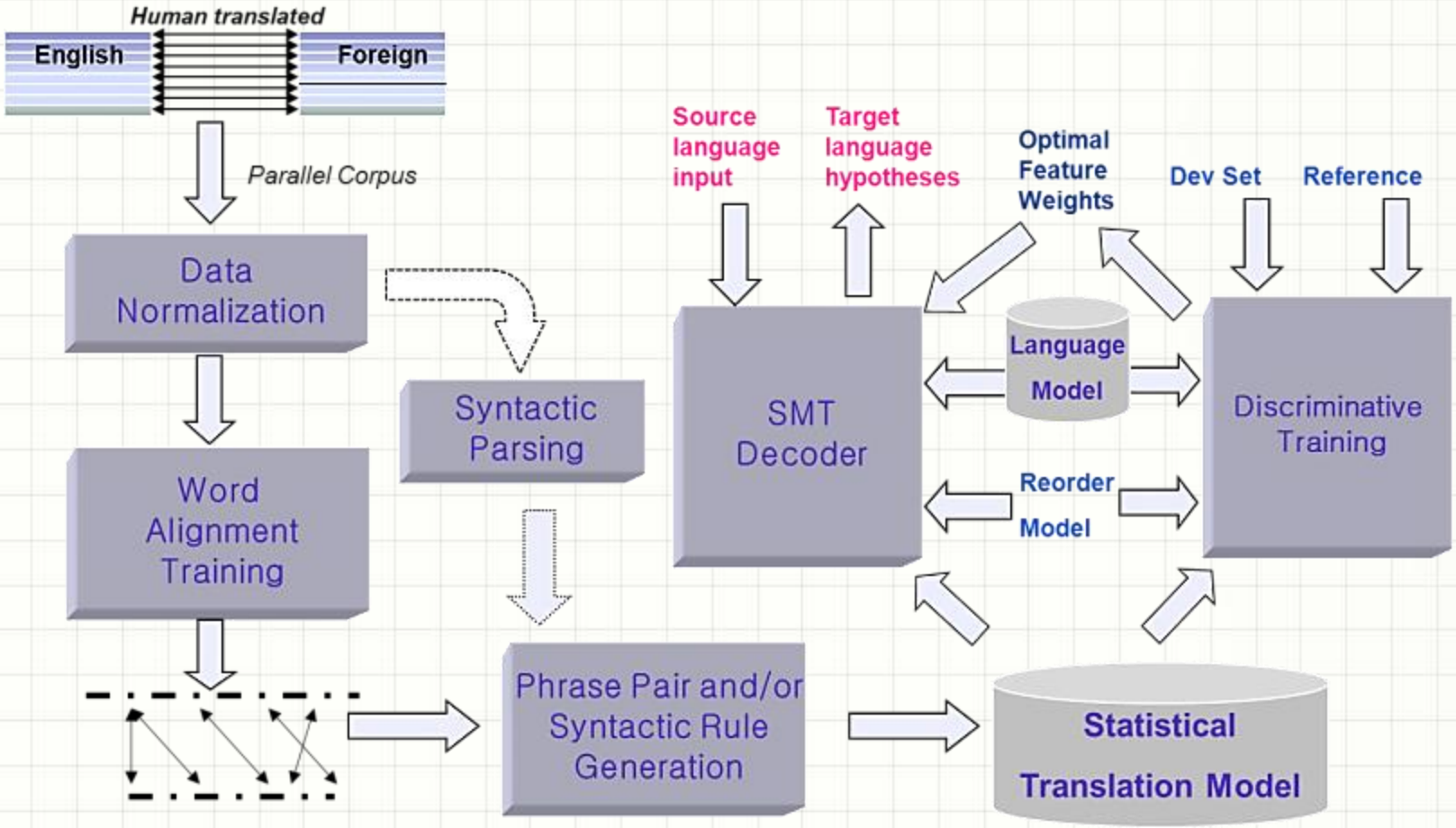
L: transduces context-independent phonetic sequences into words



A Brief History of SMT?

- MT research traced back to 1950s
- Statistical approaches started from Brown et al (1993)
  - A pioneering work based on source-channel model
  - Same approach succeeded for ASR at IBM and elsewhere
  - A good confluence example of speech and language communities to drive the transition:
    - Rationalism ➔ Empiricism (Church & Mercer, 1993; Knight, 1999)
- A sequence of unsupervised word alignment models
  - Known today as IBM Models 1-5 (Brown et al, 1993)
- Much progress has been made since then
  - Key topics for today's talk

A Bird view of SMT



How to know Translation X is better than Y?

- It's a hard problem by itself
  - More than one good translation & even more bad ones
- Assumption: "closer" to human translation(s), the better
- Metrics were proposed to measure the *closeness,* e.g., BLEU, which measures n-gram precisions (Papineni et al., 2002)

$$BLEU\text{-}4 = \mathrm{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4}\log(p_n)\right)$$

- ST measurement is more complicated
  - Objective: WER+BLEU (or METEOR, TER etc)
  - Subjective: HTER (GALE/BOLT), concept transfer rate (TransTac/BOLT)

Further Reading

- ASR Backgrounds
  - L. R. Rabiner. 1989 A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE.
  - F. Jelinek, 1997. Statistical Methods for Speech Recognition. MIT press.
  - J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan and D. O'Shaughnessy, Research developments and directions in speech recognition and understanding, IEEE Signal Processing Mag., vol. 26, pp. 75–80, May 2009.
- SMT Backgrounds
  - P. Brown, S. Pietra, V. Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, vol. 19, 1993, pp. 263-311
  - K. Knight. 1999. A Statistical MT Tutorial Workbook. 1999
- SMT Metrics:
  - K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. ACL. 2002.
  - M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proc. AMTA. 2006
  - S. Banerjee and A. Lavie. 2005 .METEOR: An automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization. 2005
  - C. Callison-Burch, P. Koehn, C. Monz, K. Peterson and M. Przybocki and O. F. Zaidan, 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In Proc. Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR
- History of MT and SMT:
  - J. Hutchins, 1986. Machine translation: past, present, future. Chichester, Ellis Horwood. Chap. 2.
  - K. W. Church and R. L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics 19.1 (1993): 1-24

Learning Problems in SMT

1. Overview: ASR, SMT, ST, Metric
2. **Learning Problems in SMT**
3. Translation Structures for ST
4. Decoding: A Unified Perspective ASR/SMT
5. Coupling ASR/SMT: Decoding & Modeling
6. Practices
7. Summary and Future Directions

Learning Translation Models

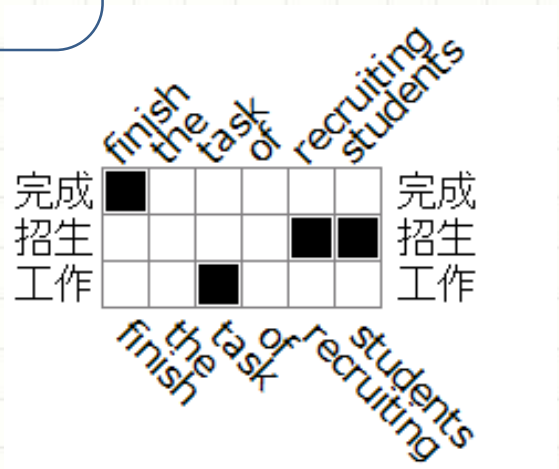
- Word alignment
- From words to phrases
- Log-linear model framework to integrate component models (a.k.a. features)
- Log-linear model training to optimize translation metrics (e.g., MERT)

Word Alignment

Given a set of parallel sentence pairs (e.g., ENU-CHS), find the word-to-word alignment between each sentence pair.

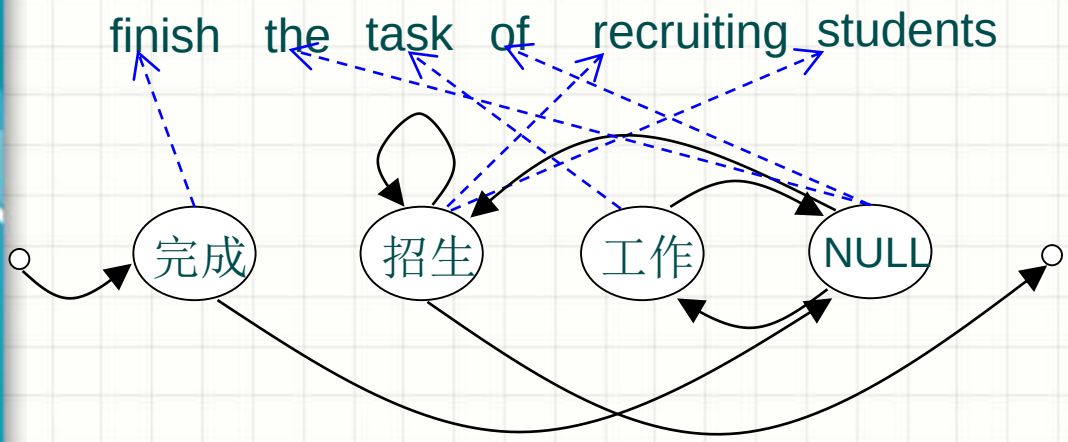
English: Finish the task of recruiting students
Chinese: 完成 招生 工作
……

word translation models



HMM for Word Alignment

Each word in the Chinese sentence is modeled as a HMM state. The English words, the *observation*, are generated by the HMM one by one.



The generative story for word alignment:

At each step, the current state (a Chinese word) emits one English word; then jump to the next state.

Similar to ASR, HMM can be used to model the word-level translation process
Unlike ASR, note the non-monotonic jumping between states, and the NULL state.

(Vogel et al., 1996)

Formal Definition

$f_1^J = f_1, \ldots, f_J$: source sentence (observation)

f_j : source word at position j

J : length of source sentence

$e_1^I = e_1, \ldots, e_I$: target sentence (states of HMM)

e_i : target word at position i

I : length of target sentence

$a_1^J = a_1, \ldots, a_J$: alignment (state sequence)

$a_j \in [1, I]: f_j \Leftrightarrow e_{a_j}$

$p(f|e)$: word translation probability

HMM Formulation

Given f_1^J and e_1^I, a_1^J is treated as "hidden variable"

Emission probability:
Model the word-to-word translation

$$p(f_1^J \mid e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} \left[p(a_j \mid a_{j-1}, I)\, p(f_j \mid e_{a_j}) \right]$$

Transition probability:
Model the distortion of the word order

Model assumption:

the emission probability only depends on the target word

the transition probability only depends on the position of the last state – *relative distortion*

ML Training of HMM

Maximum likelihood training:

$$\Lambda_{ML} = \arg\max_{\Lambda} p(f_1^J \mid e_1^I, \Lambda)$$

$$\Lambda = \left\{ p(a_j = i \mid a_{j-1} = i', I), p(f_j = f \mid e_i = e) \right\}$$

Expectation-Maximization (EM) training for Λ.

efficient Forward-Backward algorithm exists.

Find the Optimal Alignment

- Viterbi decoding:

$$\hat{a}_1^J = \arg\max_{a_1^J} \prod_{j=1}^{J} \left[p(a_j \mid a_{j-1}, I)\, p(f_j \mid e_{a_j}) \right]$$

- other variations:

  posterior probability based decoding

  max posterior mode decoding

IBM Model 1-5

- IBM model 1-5
  - A series of generative models (Brown et al., 1994)
  - Summarized below (along with the HMM)

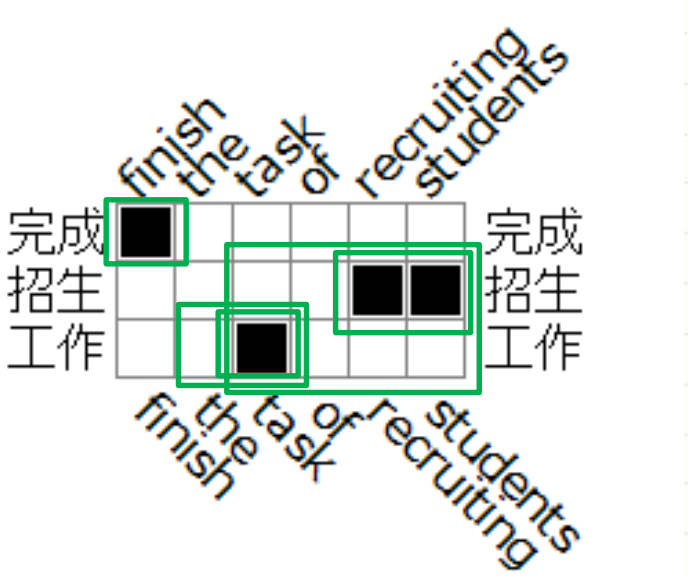
Model	property
IBM-1	Convex model (non-strictly). E-M training. Doesn't model word order distortion.
HMM	Relative distortion model (captures the intuition that words move in groups) Efficient E-M training algorithm
IBM-2	IBM-1 plus an absolute distortion model (measure the divergence of the target word's position from the ideal monotonic alignment position)
IBM-3	IBM-2 plus a fertility model (models the number of words that a state generates) Approximate parameter estimation due to the fertility model, costly training
IBM-4	Like IBM-3, but IBM-4 uses relative distortion model
IBM-5	IBM-5 addressed the model deficiency issue (while IBM-3&4 are deficient).

Other Word Alignment Models

Extensions of HMM	(Tantounova et al. 2002) (Liang et al. 2006) (Deng & Byrne 2005) (He 2007)	• Model the fertility implicitly. • Regularize the alignment by agreements of two directions. • Better distortion model.
Discriminative models	(Moore et al. 2006) (Taskar et al. 2005)	• Large linear model with many features. • Discriminative learning based on annotation
Syntax-driven models	(Zhang & Gildea 2005) (Fossum et al. 2008) (Haghighi et al. 2009)	• Use syntactic features, and/or syntactically structured models

From Word to Phrase Translation

- Extract phrase translation pairs from word alignment



Source phrase	Target phrase	Feature h …
完成	finish	…
招生	recruiting students	
工作	task	
工作	the task	
招生 工作	task of recruiting students	
…		

(Och & Ney 2004; Koehn, Och and Marcu, 2003)

Common Phrase Translation Models

Model name	Parameterization (decomposed form)	Scoring at the sentence level
Forward phrase translation model	$p(\tilde{e}\|\tilde{f}) = \frac{\#(\tilde{e},\tilde{f}) - d}{\#(*,\tilde{f})}$	$h_{fp} = \prod_k p(\tilde{e}_k\|\tilde{f}_k)$
Forward lexical translation model	$p(e\|f) = \frac{\gamma(e,f)}{\gamma(*,f)}$	$h_{fl} = \prod_k \prod_m \sum_n p(e_{k,m}\|f_{k,n})$
Backward phrase translation model	Reverse version of the forward counterpart	$h_{bp} = \prod_k p(\tilde{f}_k\|\tilde{e}_k)$
Backward lexical translation model	Reverse version of the forward counterpart	$h_{bl} = \prod_k \prod_n \sum_m p(f_{k,n}\|e_{k,m})$

Integration of All Component Models

- Use a log-linear model to integrate all component models (also known as features)

$$P(E|F) = \frac{1}{Z} exp\left\{\sum_i \lambda_i h_i(E,F)\right\}$$

$\{h_i\}$: features

$\{\lambda_i\}$: feature weights

  - Common features include:
    - Phrase translation models (e.g., the four models discussed before)
    - One or more language models in the target language
    - Counts of words and phrases
    - Distortion model – models word ordering
  - All features need to be decomposable to a word/n-gram/phrase level
- Select the translation by the best integrated score

$$\hat{E} = \underset{E}{\text{argmax}}\, P(E|F) = \underset{E}{\text{argmax}} \sum_i \lambda_i h_i(E,F)$$

(Och & Ney 2002)

Training Feature Weights

- Let's denote by $\boldsymbol{\lambda}$ as $\{\lambda_i\}$, $\boldsymbol{\lambda}$ is trained to optimize translation performance

$$\hat{\boldsymbol{\lambda}} = \underset{\{\boldsymbol{\lambda}\}}{\operatorname{argmax}} BLEU\left(\hat{E}(\boldsymbol{\lambda}, F), E^*\right)$$

$$= \underset{\{\boldsymbol{\lambda}\}}{\operatorname{argmax}} BLEU\left(\underset{E}{\operatorname{argmax}} \sum_i \lambda_i h_i(E, F), E^*\right)$$

Non-convex problem!

(Och 2003)

Minimum Error Rate Training

- Given a n-best list {*E*}, find the best $\boldsymbol{\lambda}$
  - E.g., the top scored *E* gives the best BLEU.

n-best	h_1	h_2	h_3	BLEU
E_1	$v_{1,1}$	$v_{1,2}$	$v_{1,3}$	B_1
E_2	$v_{2,1}$	$v_{2,2}$	$v_{2,3}$	B_2
E_3	$v_{3,1}$	$v_{3,2}$	$v_{3,3}$	B_3
…	…			

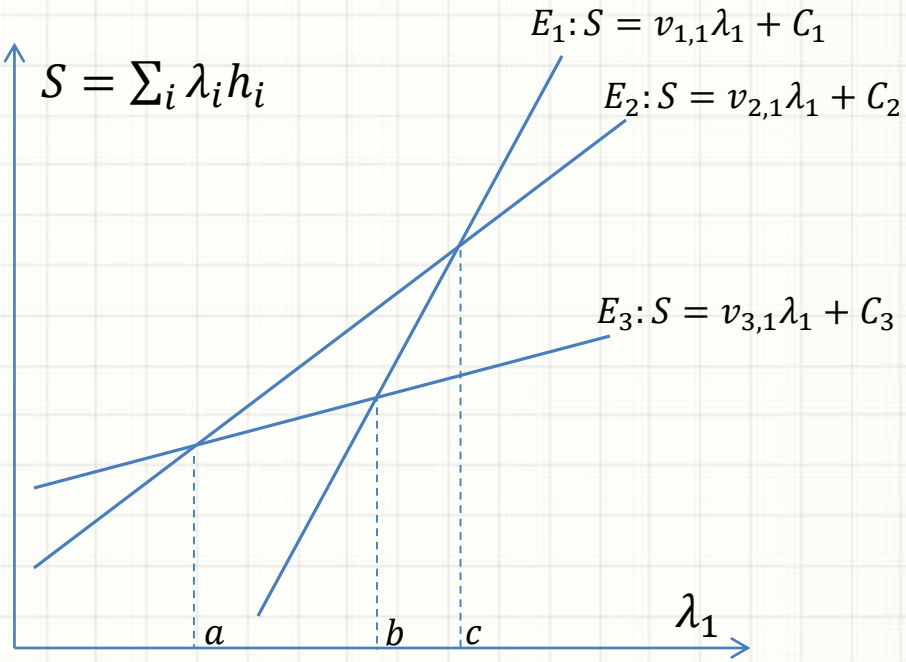


Illustration of the MERT algorithm

MERT (Och 2003):

The score of E is linear to the value of the λ that is of interest.

Only need to check several key values of λ where lines intersect with each other, since the ranking of hypotheses does not change between the two adjunct key values:

$$e.g., when\ \lambda_1 < a, E_3\ is\ top\ scored; when\ a < \lambda_1 < c, E_2\ is\ top\ scored; \ldots$$

More Advanced Training

- Use a large set of (sparse) features
  - E.g., integrate lexical, POS, syntax features
  - MERT is not effective anymore, use MIRA, PRO etc. for training of feature weights (Watanabe et al., 2007, Chiang et al. 2009 , Hopkins & May 2011, Simianer et al. 2012)
- Better estimation of (dense) translation models
  - E-M style phrase translation model training (Wuebker et al. 2010, Huang & Zhou 2009)
  - Train the translation probability distributions discriminatively (He & Deng 2012, Setiawan & Zhou 2013)
- A mix of the above two approaches
  - Build a set of new features for each phrase pair, and train them discriminatively by max expected BLEU (Gao & He 2013)

Further Reading

- P. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer. 1994. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics.
- D. Chiang, K. Knight and W. Wang, 2009. 11,001 new features for statistical machine translation. In Proceedings of NAACL-HLT.
- Y. Deng and W. Byrne, 2005, HMM Word and Phrase Alignment For Statistical Machine Translation, in Proceedings of HLT/EMNLP.
- V. Fossum, K. Knight and S. Abney. 2008. Using Syntax to Improve Word Alignment Precision for Syntax-Based Machine Translation. In Proceedings of the WMT.
- J. Gao and X. He, 2013, Training MRF-Based Phrase Translation Models using Gradient Ascent, in Proceedings of NAACL
- A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. 2009. Better word alignments with supervised ITG models. In Proceedings of ACL.
- X. He and L. Deng, 2012, Maximum Expected BLEU Training of Phrase and Lexicon Translation Models , in Proceedings of ACL
- H. Hopkins, and J. May. 2011. Tuning as ranking. In Proceedings of EMNLP.
- S. Huang and B. Zhou. 2009. An EM algorithm for SCFG in formal syntax-based translation. In Proeedings of ICASSP
- P. Koehn, F. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In Proceedings of HLT-NAACL.
- P. Liang, B. Taskar, and D. Klein, 2006, Alignment by Agreement, in Proceedings of NAACL.
- R. Moore, W. Yih and A. Bode, 2006, Improved Discriminative Bilingual Word Alignment, In Proceedings of COLING/ACL.
- F. Och and H. Ney. 2002. Discriminative training and Maximum Entropy Models for Statistical Machine Translation, In Proceedings of ACL.
- F. Och, 2003, Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of ACL.
- H. Setiawan and B. Zhou. 2013. Discriminative Training of 150 Million Translation Parameters and Its Application to Pruning, NAACL
- P. Simianer, S. Riezler, and C. Dyer, 2012. Joint feature selection in distributed stochastic learning for large-scale discriminative training in SMT. In Proceedings of ACL
- K. Toutanova, H. T. Ilhan, and C. D. Manning. 2002. Extensions to HMM-based Statistical Word Alignment Models. In Proceedings of EMNLP.
- S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based Word Alignment In Statistical Translation. In Proceedings of COLING.
- T. Watanabe, J. Suzuki, H. Tsukuda, and H. Isozaki. 2007. Online large-margin training for statistical machine translation. In Proc. EMNLP 2007.
- J. Wuebker, A. Mauser and H. Ney. 2010. Training phrase translation models with leaving-one-out, In Proceedings of ACL.
- H. Zhang and D. Gildea, 2005, Stochastic Lexicalized Inversion Transduction Grammar for Alignment, In Proceedings of ACL.

Translation Structures for ST

1. Overview: ASR, SMT, ST, Metric
2. Learning Problems in SMT
3. **Translation Structures for ST**
4. Decoding: A Unified Perspective ASR/SMT
5. Coupling ASR/SMT: Decoding & Modeling
6. Practices
7. Future Directions

Translation Equivalents (TE)

- Usually represented by some synchronous (source and target) grammar.
- The grammar choices limited by two factors.
  - *Expressiveness:* is it adequate to model linguistic equivalence between natural language pairs?
  - *Computational complexity:* is it practical to build machine translation solutions upon it?
- Need to balance both, particularly for ST,
  - Robust to informal spoken language
  - Critical speed requirement due to ST's interactive nature
  - Additional bonus if it permits effectively and efficiently integration with ASR

初步 试验	initial experiments
的 成功	the success of

$$\text{初步 } X_1 \leftrightarrow \text{initial } X_1$$
$$X_1 \text{ 的 } X_2 \leftrightarrow \text{The } X_2 \text{ of } X_1$$

Two Dominating Categories of TEs

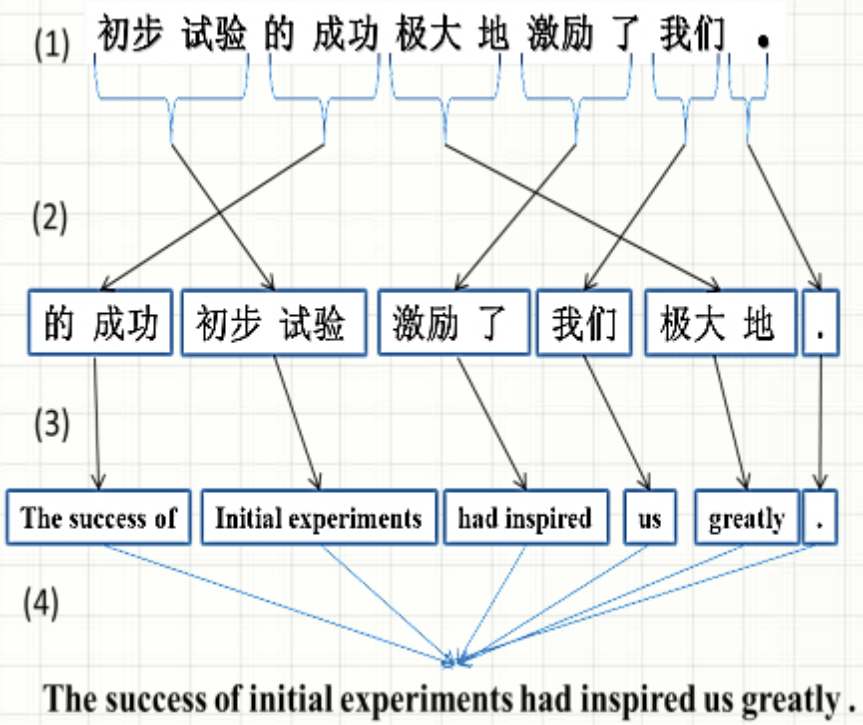
初步 试验	initial experiments
的 成功	the success of

- Finite-state-based formalism
  - E.g., phrase-based SMT (Och and Ney, 2004; Koehn et al., 2003)
- Synchronous context-free grammar-based formalism
  - hierarchical phrase-based (Chiang, 2007),
  - tree-to-string (e.g., Quirk et al., 2005; Huang et al., 2006),
  - forest-to-string (Mi et al., 2008)
  - string-to-tree (e.g., Galley et al., 04; Shen et al., 08)
  - tree-to-tree (e.g., Eisner 2003; Cowan et al, 2006; Zhang et al., 2008; Chiang 2010).
- Exceptions:
  - STAG (DeNeefe and Knight, 2009)
  - Tree Sequences (Zhang et al., 2008)

$$初步\ X_1 \leftrightarrow \text{initial}\ X_1$$
$$X_1\ 的\ X_2 \leftrightarrow \text{The}\ X_2\ \text{of}\ X_1$$

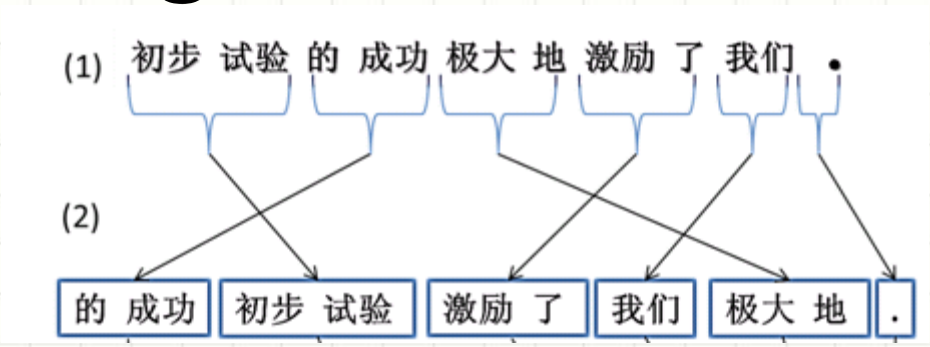
Phrase-based SMT: a generative process

1. The input sentence f_1^J is segmented into K phrases
2. Permute the source phrases in the appropriate order
3. Translate each source phrase into a target phrase
4. Concatenate target phrases to form the target sentence
5. Score the target sentence by the target language model



- *Expressiveness:* Any sequence of translation possible if permit arbitrary reordering
- *Complexity:* arbitrary reordering leads to a NP-hard problem (Knight, 1999).

Reordering in Phrasal SMT



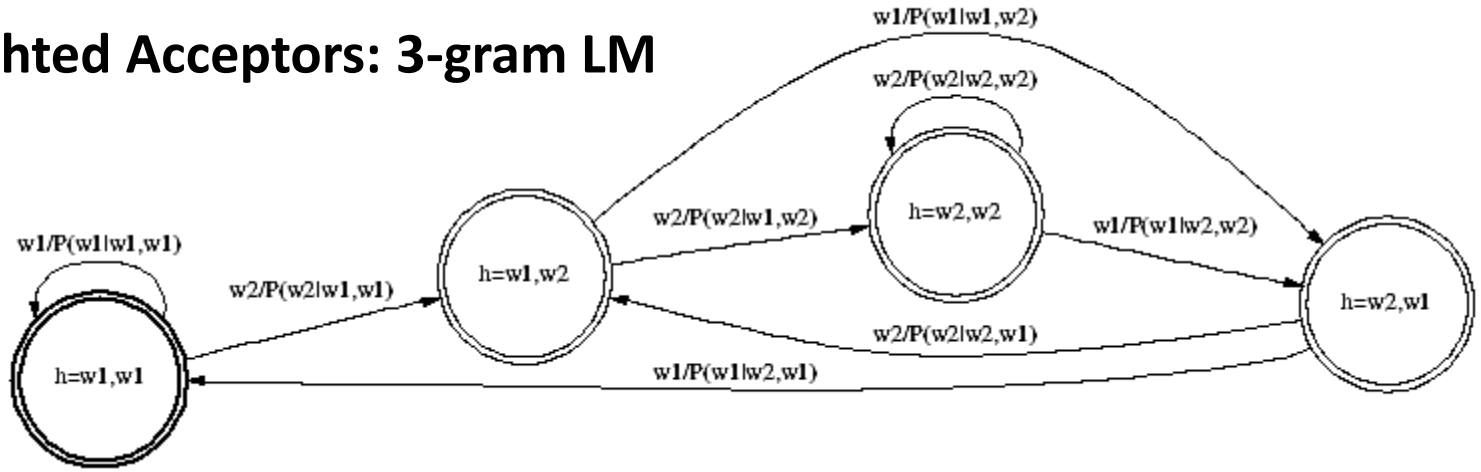
- To constrain search space, reordering is usually limited by preset maximum reordering window and skip size (Zens et al., 2004)
- Reordering models usually penalize non-monotonic movement
  - Proportional to the distance of jumping
  - Further refined by lexicalization, i.e., the degrees of penalization vary for different surrounding words (Koehn et al. 2007)
- Reordering is the unit movement at phrasal level
  - i.e., no “gaps” allowed in reordering movement

Backgrounds: Semiring

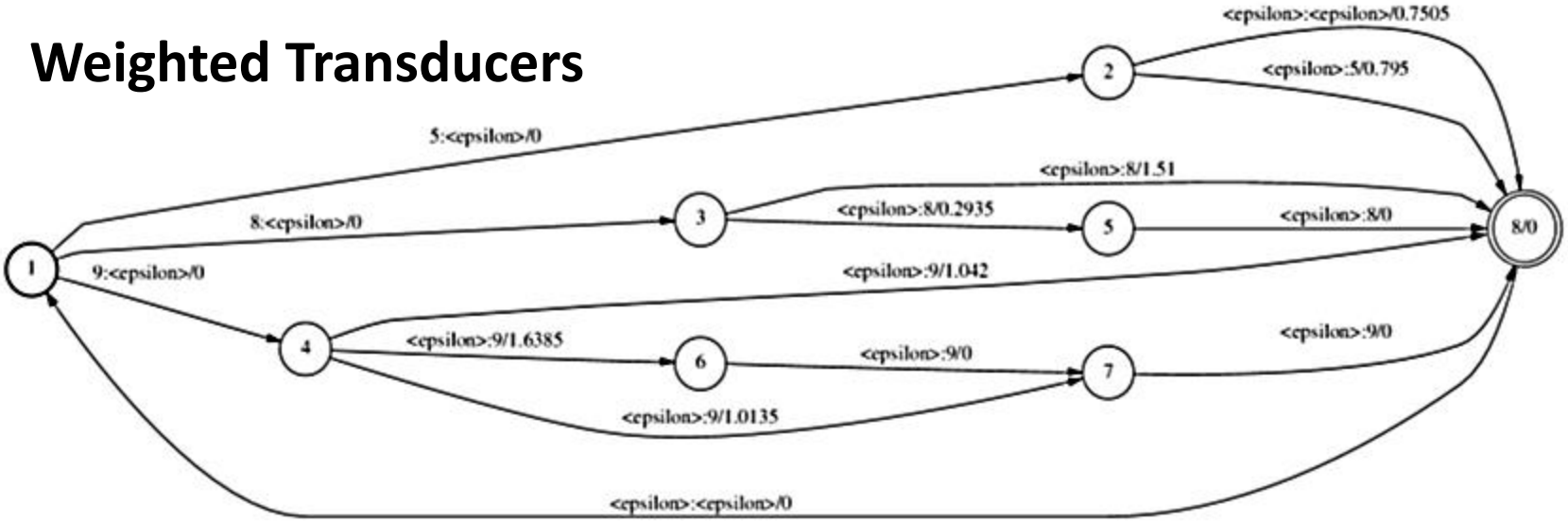
- Semiring is a 5-tuple: $(\Psi, \oplus, \otimes, \bar{0}, \bar{1})$(Mohri, 2002)
  - Ψ is a Set, and $\bar{0}, \bar{1} \in \Psi$
  - Two closed and associative operators:
    - $\oplus$ sum: to compute the weight of a sequence of edges
      - $\bar{0} \oplus a = a$ (unit); $\bar{1} \oplus a = \bar{1}$ (absorbing) $\quad \forall a \in \Psi$
    - $\otimes$ product: to compute the weight of an (optimal) path
      - $\bar{0} \otimes a = \bar{0}$ (absorbing); $\bar{1} \oplus a = a$ (unit)
- Examples used in speech & language
  - Viterbi$([0,1], \max, \times, 0, 1)$
    defined over probabilities,
  - Tropical $(\mathbf{R}^+ \cup \{+\infty\}, \min, +, +\infty, 0)$
    - operates on non-negative weights
    - e.g., negative log probabilities, aka, costs

Backgrounds: Automata/WFST

Weighted Acceptors: 3-gram LM



Weighted Transducers

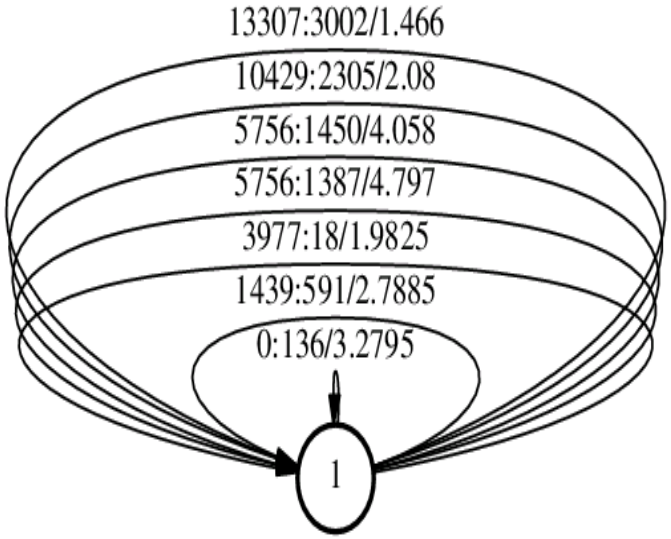


Backgrounds: Formal Languages

- A finite-state automaton (FSA) is equivalent to a regular grammar, and a regular language
- Weighted finite-state machines closed under the composition operation
- A regular grammar constitutes strict subset of a context-free grammar (CFG) (Hopcroft and Ullman, 1979)
- The composition of CFG with a FSM, is guaranteed to be a CFG

FST-based translation equivalence

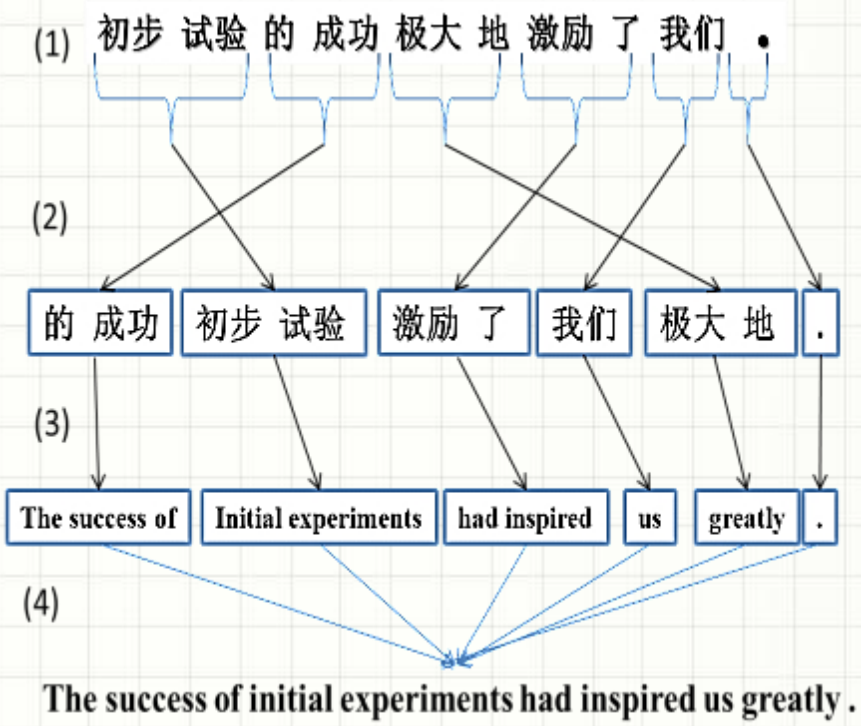
- Source-target equivalence is modeled by weighted non-nested mapping between word strings.
- Unit of mapping, $(\bar{e}_i, \bar{f}_i)$, is a modeling choice, and *phrase* is better than word due to:
  - reduced word sense ambiguities with surrounding contexts
  - appropriate local reordering encoded in the source-target phrase pair



src phrase id : tgt phrase id/cost

Phrase-based SMT is Finite-State

- Each step relates input and output as WFST operations
  - $\boldsymbol{F}$ input sentence
  - $\boldsymbol{P}$ segmentation
  - $\boldsymbol{R}$ permutation
  - $\boldsymbol{T}$ translation (phrasal equiv.)
  - $\boldsymbol{W}$ concatenation
  - $\boldsymbol{G}$ target language model
- Doing all steps amounts to composing above WFSTs
- Guaranteed to produce a FSM, since WFSTs are closed under such composition.



Similar to ASR, phrase-based SMT amounts to the following operations, computable with in a general-purpose FST toolkit (Kumar et al. 2005)

$$\hat{E} = \text{best\_path}\,(\boldsymbol{F} \circ \boldsymbol{P} \circ \boldsymbol{R} \circ \boldsymbol{T} \circ \boldsymbol{W} \circ \boldsymbol{G})$$

Why do we care?

- Not only of theoretic convenience to conceptually connect ASR and SMT better
- Practically useful to leverage the mature WFST optimization algorithms.
- Suggests integrated framework to address ST:
  - i.e., composing the WFSTs from ASR/SMT

Make WFST Approach Practical

- WFST-based system in (Kumar et al., 2005) runs significantly slower than the multiple-stack based decoder (Koehn, 2004)
  - large memory requirements
  - heavy online computation for each composition.
- Reordering is a big challenge
  - The number of states needed is $O(2^J)$; cannot be bounded for any arbitrary inputs as finite-state
- Next, we show a framework to address both issues to make it more practically suitable for ST

Folsom: phrase-based SMT by FSM

- To speed up: first cut the number of online compositions

$$\hat{E} = \text{best\_path}\,(\boldsymbol{F} \circ \boldsymbol{P} \circ \boldsymbol{R} \circ \boldsymbol{T} \circ \boldsymbol{W} \circ \boldsymbol{G}) \quad (1)$$

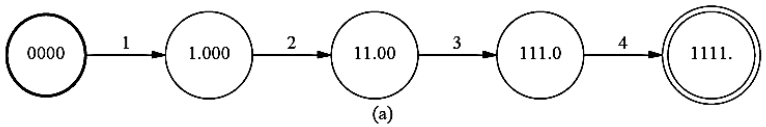
$$\hat{E} = \text{best\_path}\,(\boldsymbol{F}' \circ \boldsymbol{M} \circ \boldsymbol{G}) \quad (2)$$

  - $\boldsymbol{M}$: WFST encoding a log-linear phrasal translation model, obtained by

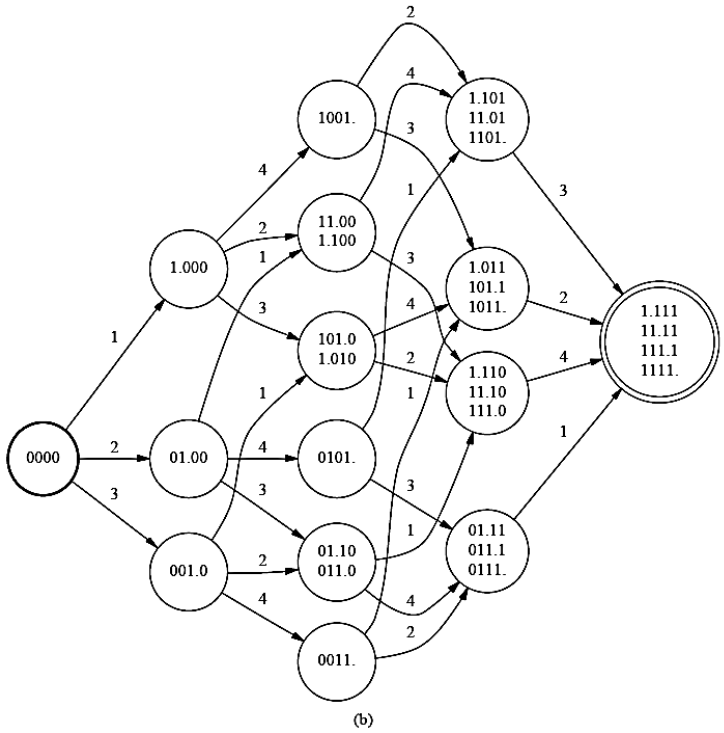
$$\boldsymbol{M} = \boldsymbol{Min}(\boldsymbol{Min}(\boldsymbol{Det}(\boldsymbol{P}) \circ \boldsymbol{T}) \circ \boldsymbol{W}$$

  - Possible by making $\boldsymbol{P}$ determinizable (Zhou et al., 2006)
- More flexible reordering: $\boldsymbol{F}'$is a WFSA constructed on-the-fly
  - To encode the uncertainty of the input sentence (e.g., reordering); examples on the next Slide
- Design a dedicated decoder for further efficiency: Viterbi search on the *lazy 3-way* composed graph as in (2)

Reordering, Finite-State & Gaps



(a): Monotonic (like ASR)



(b): Reordering with skip less than three

- Each state denotes a specific input coverage with a bit-vector.
- Arcs labeled with an input position to be covered in state transition: weights given by reordering models.
- Source segmentation allows options with "*gaps*" in searching for the best-path to transduce the inputs at the phrase level

$$f_1 f_2 f_3 f_4 \xrightarrow{\text{phrase segmentation}} \overline{f_1 f_4}\ \ \overline{f_2 f_3}$$

- Such a translation phenomenon is beneficial in many language pairs.
  - translate disjoint Chinese preposition "在…外" as "*outside*"
  - verb phrases, "唱…歌" as "*sing*".
- Reordering with gaps adds flexibility to conventional phrase-based models and confirmed to be useful in many studies

Every path from the initial to the final state represents an acceptable input permutation

What FST-based models lack

- FST-based models (e.g., phrase-based) remain state-of-the-art for many language pairs/tasks (Zollmann et al., 2008)
- It makes no use that natural language is inherently structural & hierarchical
  - Led to poor long-distance reordering modeling & exponential complexity of permutation
- PCFGs effectively used to model linguistic structures in many monolingual tasks (e.g., parsing)
  - Recall: RL is a subset of context-free language
- The synchronous (bilingual) version of the PCFG (i.e., SCFG) is a alternative to model translation structures

SCFG-based TE

- A SCFG (Lewis and Stearns, 1968) rewrites the non-terminal (NT) on its left-hand side (LHS) into <source, target> on its right-hand side (RHS)
  - s.t. the constraint of one-to-one correspondences (co-indexed by subscripts) between the source and target of every NT occurrence.

$$\mathrm{VP} \xrightarrow{p} < \mathrm{PP}_1 \mathrm{VP}_2 \, , \mathrm{VP}_2 \mathrm{PP}_1 >$$

- NTs on RHS can be recursively instantiated simultaneously for both the source and the target, by applying any rules with a matched NT on the LHS.
- SCFG captures the hierarchical structure of NL & provides a more principled way to model reordering
  - e.g., the PP and VP will be reordered regardless of their span lengths

Expressiveness vs. Complexity

- Both depend on maximum number of NTs on RHS (aka, rank of the SCFG grammar)
- Complexity is polynomial: higher order for higher rank
  - Cubic $O(|J|^3)$ for rank-two (binary) SCFG
- Expressiveness: more expressive with higher rank; Specifically for a binary SCFG
  - Rare reordering examples exist (Wu 1997; Wellington et al., 2006) that it cannot cover
  - Arguably sufficient in practice

Pic from http://translation-blog.trustedtranslations.com/quality-and-speed-in-machine-translation-2011-12-30.html

What's in common? SCFG-MT vs.

- Like ice-cream, SCFG models come with different flavors…
- Linguistic syntax based:
  - Utilize structures defined over linguistic theory and annotations (e.g., Penn Treebank)
  - SCFG rules are derived from the parallel corpus guided by explicitly parsing on at least one side of the parallel corpus.
  - E.g., tree-to-string (e.g., Quirk et al., 2005; Huang et al., 2006), forest-to-string (Mi et al., 2008) string-to-tree (e.g., Galley et al., 04; Shen et al., 08) tree-to-tree (e.g., Eisner 2003; Cowan et al, 2006; Zhang et al., 2008; Chiang 2010).
- Formal syntax based:
  - Utilize the hierarchical structure of natural language only
  - Grammars extracted from the parallel corpus without using any linguistic knowledge or annotations.
  - E.g., ITG (Wu, 1997) and hierarchical models (Chiang, 2007)

Formal syntax based SCFG

- Arguably a better for ST: not relying on parsing, which might be difficult for informal spoken languages

- Grammars: only one universal NT X is used (Chiang, 2007)
  - Phrasal rules:

$$X \rightarrow < \text{初步 试验} , \text{initial experiments} >$$

  - Abstract rules: with NTs on RHS

$$X \rightarrow < X_1 \text{ 的 } X_2 , \text{The } X_2 \text{ of } X_1 >$$

  - Glue rules to generate sentence symbol S

$$S \rightarrow < S_1 X_2 , S_1 X_2 >$$

$$S \rightarrow < X_1 , X_1 >$$

Learning of (Abstract) SCFG Rules

- Hierarchical SCFG rule extraction (Chiang 2007):
  - Replace any aligned sub-phrase pair of a PP with co-indexed NT symbols
  - Many-to-many mapping between phrase pairs and derived abstract rules.
  - Conditional probabilities estimated based on heuristic counts.
- Linguistic SCFG rule extraction
  - Syntactic parsing on at least one side of the parallel corpus.
  - Rules are extracted along with the parsing structures: constituency (Galley et al., 2004; Galley et al., 2006), and dependency (Shen et al., 2008)
- Improved extraction and parameterization:
  - Expected counts: forced alignment and inside-outside (Huang and Zhou, 2009);
  - Leaving-one-out smoothing (Wuebker et al., 2010)
  - Extract additional rules:
    - Reduce conflicts between alignment or parsing structures (DeNeefe et al., 2007)
    - From existing ones of high expected counts: rule arithmetic (Cmejrek and Zhou, 2010)

Practical Considerations of Formal Syntax-based SCFG

- On Speed:
  - Finite-state based search with limited reordering usually runs faster than most of the SCFG-based models.
  - Except for tree-to-string (e.g., Huang and Mi, 2010), which is faster in practice.
- On Model:
  - With beam pruning, using higher-rank (>2) SCFG is possible
  - Weakness: no constraints on X often led to over-generalization of SCFG rules
- Improvements: add linguistically motivated constraints
  - Refined NT with direct linguistic annotations (Zollmann and Venugopal, 2006)
  - Soft constraints (Marton et al., 2008)
  - Enriched features tied to NTs (Zhou et al., 2008b, Huang et al., 2010)

Further Reading

- Phrase-based SMT
  - F. Och and H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics. 2004
  - F. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, vol. 29, 2003, pp. 19-51
  - F. Och and H. Ney. 2002. Discriminative training and Maximum Entropy Models for Statistical Machine Translation, In Proceedings of ACL.
  - P. Koehn, F. J. Och and D. Marcu. 2003. Statistical phrase based translation. In Proc. NAACL. pp. 48 – 54.
  - P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL Demo Paper. 2007.
  - S. Kumar, Y. Deng and W. Byrne, 2005. A weighted finite state transducer translation template model for statistical machine translation, Journal of Natural Language Engineering, vol. 11, no. 3, 2005. Proc. ACL-HLT. 2008
  - R. Zens, H. Ney, T. Watanabe and E. Sumita. 2004. Reordering constraints for phrase-based statistical machine translation," In Proc. COLING. 2004
  - B. Zhou, S. Chen and Y. Gao. 2006. Folsom: A Fast and Memory-Efficient Phrase-based Approach to Statistical Machine Translation. In Proc. SLT. 2006
- Computational theory:
  - J. Hopcroft and J. Ullman. 1979. Introduction to Automata Theory, Languages, and Computation, Addison-Wesley. 1979.
  - G. Gallo, G. Longo, S. Pallottino and S. Nyugen, 1993. Directed hypergraphs and applications. Discrete Applied Mathematics, 42(2).
  - M. Mohri. 2002. Semiring frameworks and algorithms for shortest-distance problems," Journal of Automata, Languages and Combinatorics, vol. 7, no. 3, p. 321–350, 2002.
  - M. Mohri, F. Pereira and M. Riley Weighted finite-state transducers in speech recognition. Computer Speech and Language, vol. 16, no. 1, pp. 69-88, 2002.

Further Reading: SCFG-based SMT

- B. Cowan, I. Ku˘cerov´a and M. Collins. 2006. A discriminative model for tree-to-tree translation. In Proc. EMNLP 2006.
- D. Chiang. 2007. Hierarchical phrase-based translation. Computational Linguistics, 33(2):201–228.
- D. Chiang, K. Knight and W. Wang, 2009. 11,001 new features for statistical machine translation. In Proc. NAACL-HLT, 2009.
- D. Chiang. 2010. Learning to translate with source and target syntax. In Proc. ACL. 2010.
- S. DeNeefe, K. Knight, W. Wang and D. Marcu. 2007. What Can Syntax-Based MT Learn from Phrase-Based MT? In Proc. EMNLP-CoNLL. 2007
- S. DeNeefe and K. Knight, 2009. Synchronous tree adjoining machine translation. In Proc. EMNLP 2009
- J. Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In Proc. ACL 2003
- M. Galley and C. Manning. 2010. Accurate nonhierarchical phrase-based translation. In Proc. NAACL 2010.
- M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang and I. Thayer, 2006. Scalable inference and training of context-rich syntactic translation models. In Proc. COLING-ACL. 2006.
- M. Galley, M. Hopkins, K. Knight and D. Marcu. 2004. What's in a translation rule? In Proc. NAACL. 2004.
- L. Huang and D. Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In Proc. ACL. 2007
- G. Iglesias, C. Allauzen, W. Byrne, A. d. Gispert and M. Riley. 2011. Hierarchical Phrase-Based Translation Representations. In Proc. EMNLP
- G. Iglesias, A. d. Gispert, E. R. Banga and W. Byrne. 2009. Hierarchical phrase-based Translation with weighted finite state transducers. In Proc. NAACL-HLT. 2009
- H. Mi, L. Huang and Q. Liu. 2008. Forest-based translation. in Porc. ACL 2008
- C. Quirk, A. Menezes and C. Cherry. 2005. Dependency treelet translation: Syntactically informed phrase-based SMT. In Proc. ACL.
- L. Shen, J. Xu and R. Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In Proc. ACL-HLT. 2008
- D. Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics. vol. 23, no. 3, pp. 377-403, 1997.
- M. Zhang, H. Jiang, A. Aw, H. Li, C. L. Tan and S. Li. 2008. A Tree Sequence Alignment-based Tree-to-Tree Translation Model. In Proc. ACL-HLT. 2008

Decoding: A Unified Perspective for ASR/SMT/ST

1. Overview: ASR, SMT, ST, Metric
2. Learning Problems in SMT
3. Translation Structures for ST
4. **Decoding: A Unified Perspective ASR/SMT**
5. Coupling ASR/SMT: Decoding & Modeling
6. Practices
7. Future Directions

Unifying ASR/SMT/ST Decoding

- Upon first glance, there is dramatic difference between ASR and SMT due to reordering
- Even for SMT, there are a variety of paradigms
  - Each may call for its own decoding algorithm
- Common in all decoding:
  - Search space is usually exponentially large and DP is a must
  - DP: divide-and-conquer w/ reusable sub-solutions.
  - Well-known Viterbi search in HMM-based ASR: an instance of DP
- We try unify them by observing from a higher standpoint
- Benefits of a unified perspective
  - Help us understand concepts better
  - Reveals a close connection between ASR and SMT
  - Beneficial for joint ST decoding

Background: Weighted Directed Acyclic Graph $\Sigma = (V, E, \Omega)$

- A vertices set V and edges set E
- A weight mapping function Ω: $E \rightarrow \Psi$ assigns each edge a weight from Ψ
  - Ψ defined in a semiring $(\Psi, \oplus, \otimes, \bar{0}, \bar{1})$
- A single source vertex $s \in V$ in the graph.
- A path π in G is a sequence of consecutive edges $\pi = e_1 e_2 \cdots e_l$
  - End vertex of one edge is the start vertex of the subsequent edge
  - Weight of path $\Omega(\pi) = \bigotimes_{i=1}^{l} \Omega(e_i)$
- The shortest distance from s to a vertex q, $\delta(q)$, is the "$\oplus$-*sum*" of the weights of all paths from s to q
- We define $\delta(s) = \bar{1}$

The search space of any finite-state automata, e.g., the one used in HMM-based ASR and FST-based translation, can be represented by such a graph.

Genetic Viterbi Search of DAG

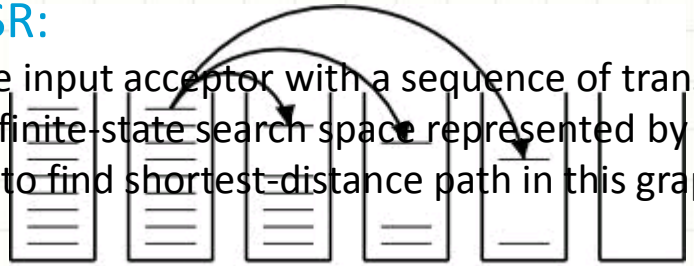
Algorithm 1: Generic Viterbi Search of DAG

1. **Procedure** Viterbi (Σ, s)
2. Initialize Σ, s
3. Topological sort vertices of Σ
4. **for** each vertex p in Σ in topological order **do**
5. **for** each edge e such that $start(e) = p$ **do**
6. $q = end(e)$
7. $\delta(q) \oplus= \delta(p) \otimes w(e)$

- Many search problems can be converted to the classical shortest distance problem in the DAG (Cormen et al, 2001)
- Complexity is $O(|V| + |E|)$, as each edge needs to be visited exactly once.
- Terminates when all reachable vertices from s have been visited, or if some predefined destination vertices encountered

Case Studies: ASR and Multi-stack Phrasal SMT

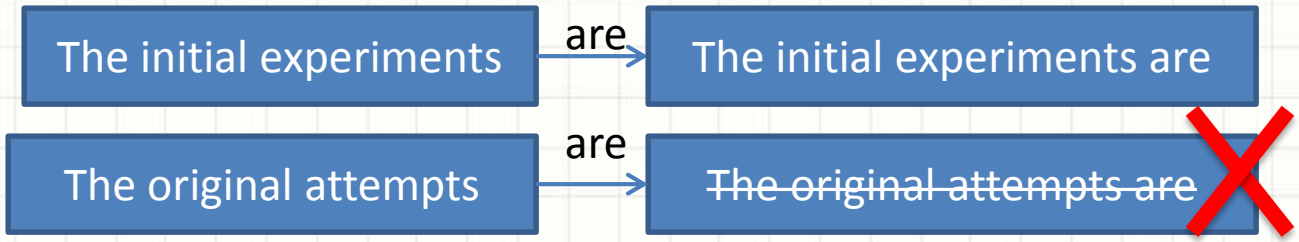
- HMM-based ASR:
  - Composing the input acceptor with a sequence of transducer (Mohri et al., 2002)
  - This defines a finite-state search space represented by a graph
  - Viterbi search to find shortest-distance path in this graph



$$\hat{f}_1^J = \text{best\_path}\,(\boldsymbol{O} \circ \boldsymbol{H} \circ \boldsymbol{C} \circ \boldsymbol{L} \circ \boldsymbol{G})$$

- Multi-stack Phrasal SMT e.g., Moses (Koehn et al., 2007)
  - The source vertex: none of the source words has been translated and the translation hypothesis is empty.
  - The target vertex: all the source words have been covered
  - Each edge connect start and end vertices by choosing a consecutive uncovered set of source words, to apply one of the source-matched phrase pairs (PP)
  - The target side of the PP is appended to the hypothesis
  - The weight of each edge is determined by a log-linear model.
  - Such edges are expanded consecutively until arrives at t
  - The best translation collected along the best path with the lowest weight $\delta(t)$

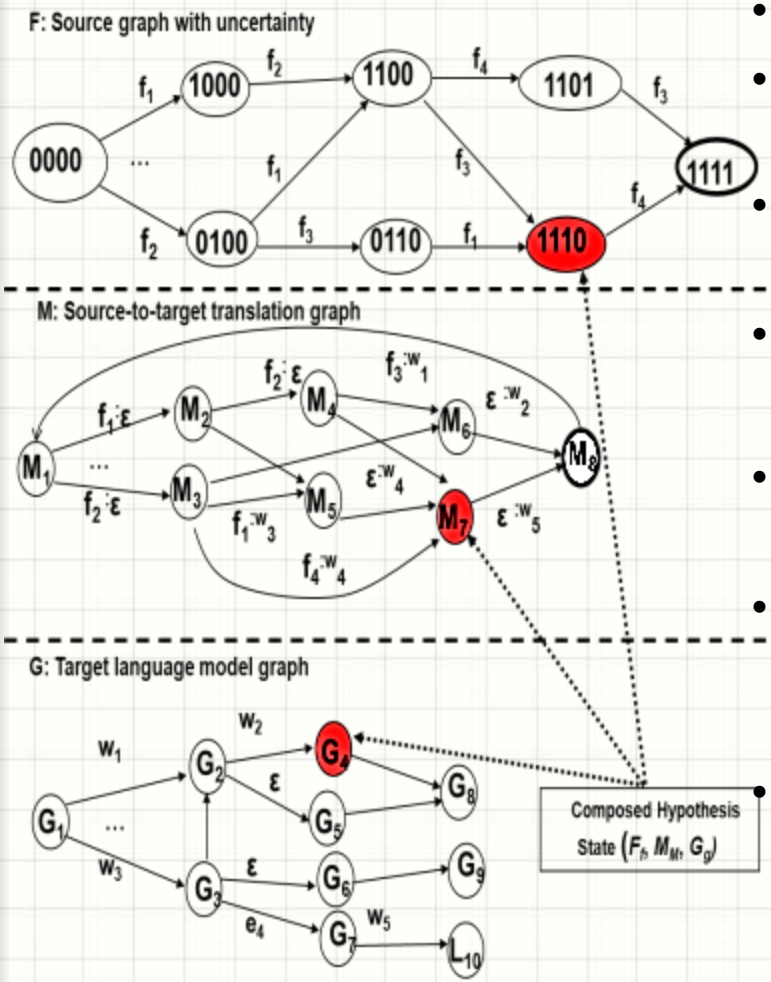
Hypothesis Recombination in Graph



- key idea: keep partial hypothesis with the lowest weights and discard others with the same *signature* in phrase-based SMT
  - *same* set of source words being covered
  - *same* n-1 last words (for n-gram LM) in the hypothesis
- Only the partial hypothesis with the lowest cost has a possibility to become the best hypothesis in the future
  - So HR is loses-free for 1-best search
- It is clear why this works if we view this from graph
  - All partial hypotheses of same signature arrive at the same vertex
  - All future paths leaving this vertex are indistinguishable afterwards.
  - Under "$\oplus$-*sum*" operation (e.g., the *min* in the Tropical semiring), only the lowest-cost partial hypothesis is kept

Case Studies II: Folsom Phrasal SMT

$\hat{E} = \text{best\_path}\ (\boldsymbol{F'} \circ \boldsymbol{M} \circ \boldsymbol{G})$



- Graph expanded by lazy 3-way composition
- Source vertex = start states comp. (F_1, M_1, G_1)
- Target vertices = each component state is a final state in each individual WFST
- Subsequent vertices visited in topological order
- Weight(e) from (F_p, M_p, G_p) to (F_q, M_q, G_q) :
$$\Omega(e) = \bigotimes_{m \in \{I,M,G\}} \lambda_m \Omega(e_m)$$
- HR: merge vertices of same (F_q, M_q, G_q)& keep only the one with lowest cost
- Any path connecting the source to a target vertex is a translation: best one with the shortest distance.

The decoding is optimized lazy *3-way* composition + minimization, followed by best-path search.

Background: Weighted Directed Hypergraph $H =< V, E, \Psi >$

- A vertices set V and hyperedge set E
- Each hyperedge$e$ links an ordered list of tail vertices to a head vertex
- *Arity* of H is the maximum number of tail vertices of all e
- $f_e: \Psi^{|T(e)|} \rightarrow \Psi$ assigns each hyperedge e a weight from Ψ
- A derivation d of a vertex q: a sequence of consecutive e connecting source vertices to q
  - Weight $\Omega(d)$ recursively computed from *weight functions* of each e.
- The "best" weight of q is the "$\oplus$-*sum*" of all of its derivations

$$\delta(q) = \begin{cases} \bar{1}, & \text{q is a source vertex} \\ \bigoplus_d \Omega(d), & \text{otherwise} \end{cases}$$

A hypergraph, a generalization of a graph, encodes the hierarchical-branching search space expanded over CFG models

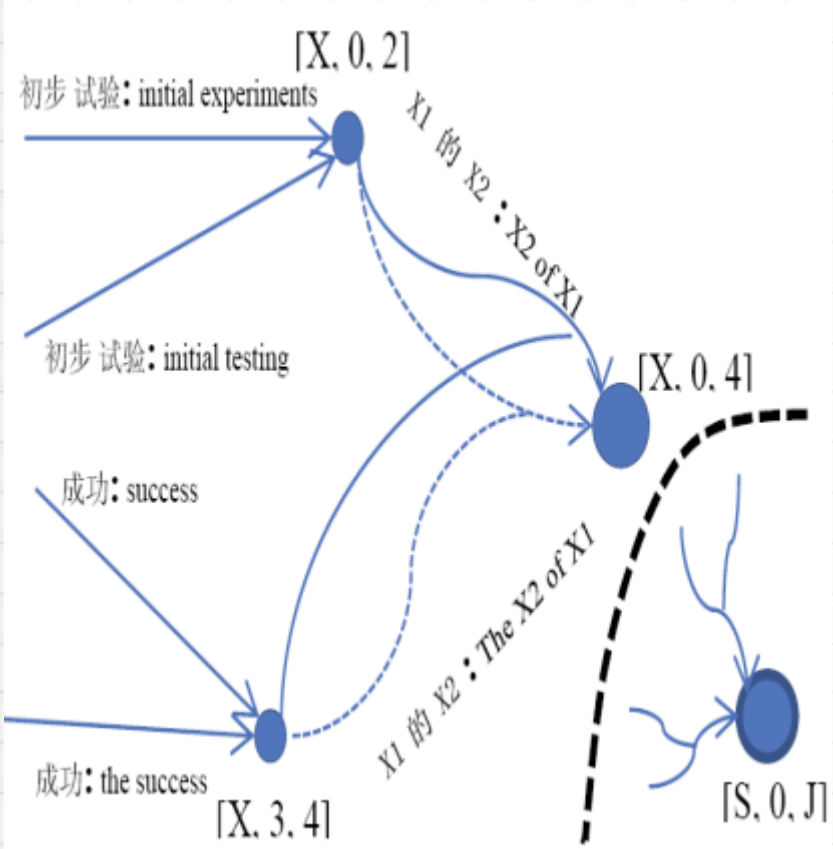
Genetic Viterbi Search of DAH

Algorithm 2: Generic Viterbi Search of DAH

1. **Procedure** Viterbi (H, s_1^n)
2. Initialize H, s_1^n
3. Topological sort the vertices of H
4. **for** each vertex q in H in topological order **do**
5. **for** each hyperedge e such that $head(e) = q$ **do**
6. $\{p_1, \cdots, p_{|T(e)|}\} = tail(e)$
7. $\delta(q) \oplus= f_e(\delta(p_1), \cdots, \delta(p_{|T(e)|}))$

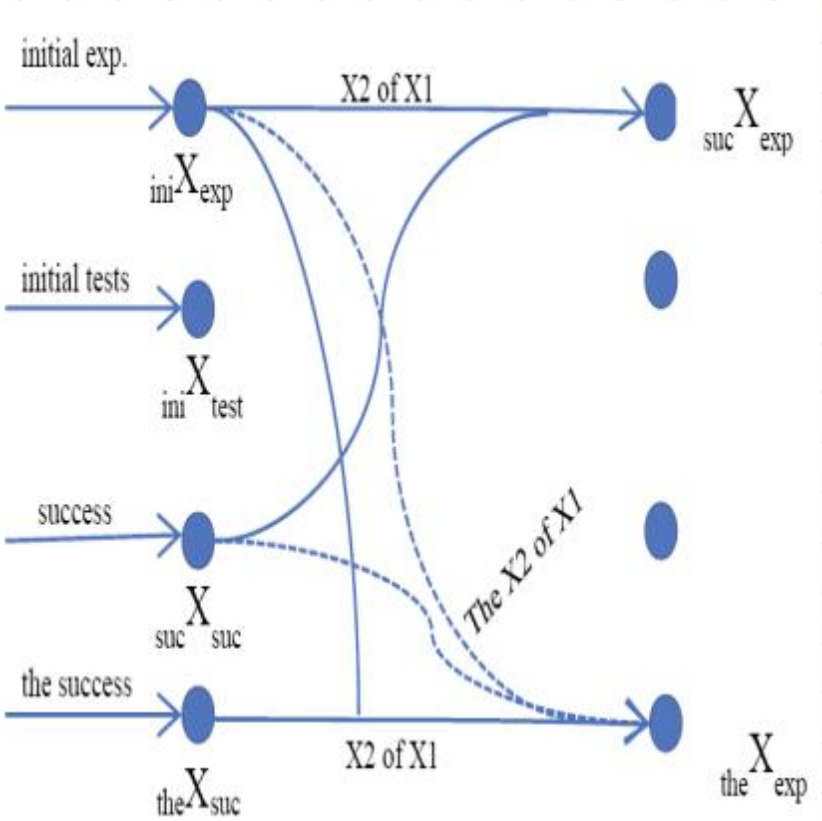
- Decoding of SCFG-based models largely follows CKY, an instance of Viterbi on DAH of arity two
- Complexity is proportional to $O(|E|)$

Case Study: Hypergraph decoding



- Vertices: $[X, i, j]$ the LHS NT + span
- Source vertices: apply phrasal rules matching any consecutive inputs
- Topological sort ensures that vertices with shorter spans $j - i$ are visited earlier than those with longer spans
- Derivation weight $\delta(q)$ at head vertex updated per line 7
- Process repeated until the target vertex $[S, 0, J]$ is reached.
- The best derivation found by back-tracing from the target vertex
- Complexity $O(|E|) \propto O(|\mathcal{R}|J^3)$

Case Study: DAH with LM



- Search space is still DAH
- Tail vertex additionally encodes the n-1 words on both the left-most and right-most side
- Each hyperedge updates LM score and boundary information.
- Worst-case complexity is $O\left(|\mathcal{R}|J^3|T|^{4(n-1)}\right)$
- Pruning is a must

Pruning: A perspective from graph

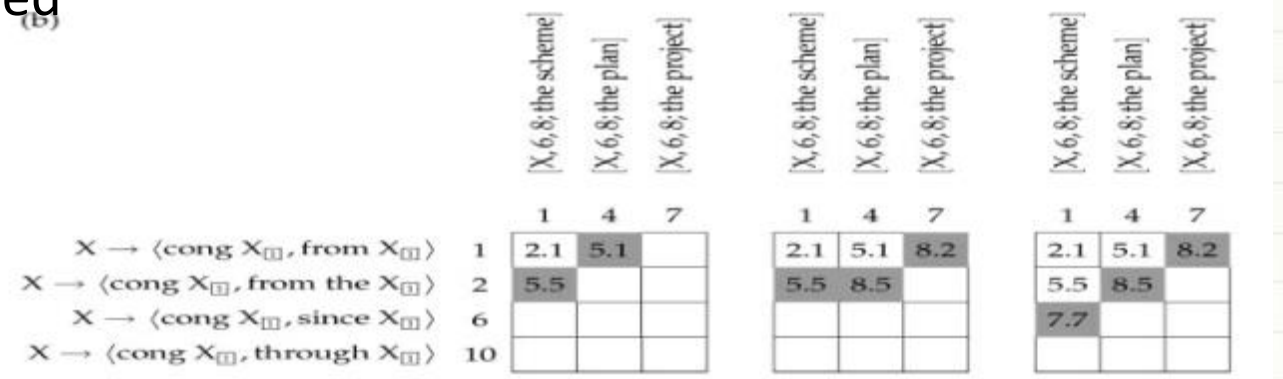
- Two options to prune in a graph

1. Discard some end vertices
   - Sort active vertices (sharing the same coverage vector) based on $\delta(q)$
   - skipping certain vertices that are outside either
     - a beam from the best (*beam pruning)*
     - the top k list (*histogram pruning*).
2. Discard some edges leaving a start vertex (*beam or histogram pruning)*
   - Apply certain reordering constraints (Zens and Ney, 2004)
   - Discard higher-cost phrase-based translation options.

- Both, unlike HR, may lead to search errors.

A lazy Histogram Pruning

- Avoid expanding edges *if* they fall outside of the top k, if
  - weight function is monotonic in each of its arguments,
  - Input list for each argument is sorted.
- Example: cube pruning for DAH search (Chiang,07):
  - Suppose a *hyperedge bundle* $\{e_j\}$ where they share the same source side and identical tail vertices
  - Hyperedges with lower costs are visited earlier
  - Push e into a priority queue that is sorted by $f_e: \Psi^{|T(e)|} \rightarrow \Psi$.
  - Hyperdge popped from the priority queue is explored.
  - Stops when the top k hyperedges popped, and all remaining ones discarded



Further Reading

- D. Chiang. 2007. Hierarchical phrase-based translation. Computational Linguistics, 33(2):201–228.
- G. Gallo, G. Longo, S. Pallottino and S. Nyugen, 1993. Directed hypergraphs and applications. Discrete Applied Mathematics, 42(2).
- L. Huang. 2008. Advanced dynamic programming in semiring and hypergraph frameworks. In Proc. COLING. Survey paper to accompany the conference tutorial. 2008
- M. Mohri, F. Pereira and M. Riley Weighted finite-state transducers in speech recognition. Computer Speech and Language, vol. 16, no. 1, pp. 69-88, 2002.
- D. Klein and C. D. Manning. 2004. Parsing and hypergraphs. New developments in parsing technology, pages 351–372.
- K. Knight. 1999. Decoding Complexity in Word-Replacement Translation Models. Computational Linguistics, vol. 25, no. 4, pp. 607-615, 1999.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL Demo Paper. 2007.
- L. Huang and D. Chiang. 2005. Better k-best parsing. In Proc. the Ninth International Workshop on Parsing Technologies. 2005
- L. Huang and D. Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In Proc. ACL. 2007
- B. Zhou. 2013. Statistical Machine Translation for Speech: A Perspective on Structure, Learning and Decoding, in *Proceedings of the IEEE*, IEEE, August 2013

Coupling ASR and SMT for ST – Joint Decoding

1. Overview: ASR, SMT, ST, Metric
2. Learning Problems in SMT
3. Translation Structures for ST
4. Decoding: A Unified Perspective ASR/SMT
5. **Coupling ASR/SMT: Decoding** & Modeling
6. Practices
7. Future Directions

Bayesian Perspective of ST

- Sum replaced by Max
- A common practice

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} P(e_1^I | x_1^T)$$

$$= \underset{e_1^I}{\operatorname{argmax}} \left\{ \sum_{f_1^J} P(e_1^I | f_1^J) P(x_1^T | f_1^J) P(f_1^J) \right\}$$

$$\approx \underset{e_1^I}{\operatorname{argmax}} \left\{ \max_{f_1^J} P(e_1^I | f_1^J) P(x_1^T | f_1^J) P(f_1^J) \right\}$$

$$\approx \underset{e_1^I}{\operatorname{argmax}} \left\{ P[e_1^I | \underset{f_1^J}{\operatorname{argmax}} P(x_1^T | f_1^J) P(f_1^J)] \right\}$$

- f_1^J determined only by x_1^T, and solved as an isolated problem.
- The foundation of the cascaded approach

- Cascaded approach impaired by compounding of errors propagated from ASR to SMT
- ST improved if ASR/SMT interactions are factored in with better coupled components.
- Coupling achievable with joint decoding and more coherent modeling across components

Tight Joint Decoding

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \left\{ \max_{f_1^J} P(e_1^I | f_1^J) P(x_1^T | f_1^J) P(f_1^J) \right\}$$

- A fully integrated search over all possible e_1^I and f_1^J
- With the unified view of ASR/SMT, it's feasible for translation models with *simplified* reordering
  - WFST-based word-level ST (Matusov et al. 2006):
    - substituted $P(e_1^I, f_1^J)$ for the usual source LM used in the ASR WFSTs
    - Produce speech translation in the target language.
  - Monotonic joint translation model at phrase-level (Casacuberta et al., 2008).
  - Tight joint decoding using phrase-based SMT can be achieved by Folsom
    - Composing ASR WFSTs with $\boldsymbol{M}$ and $\boldsymbol{G}$ on-the-fly
    - Followed by fully integrated search
    - Limitation is that reordering can only occur within a phrase.

$$\hat{E} = \text{best path}\,(\boldsymbol{X} \circ \boldsymbol{H} \circ \boldsymbol{C} \circ \boldsymbol{L} \circ \boldsymbol{M} \circ \boldsymbol{G})$$

Loose Joint Decoding

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \left\{ \max_{f_1^J} P(e_1^I | f_1^J) P(x_1^T | f_1^J) P(f_1^J) \right\}$$

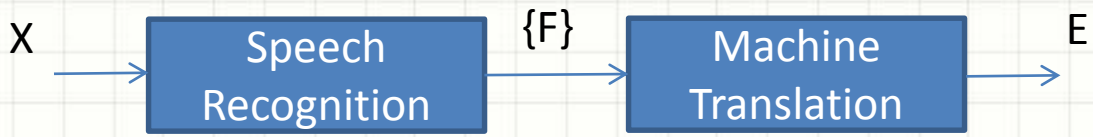
- Approximating full search space with a promising subset
  - N-best ASR hypotheses (Zhang et al., 2004): Weak
  - Word lattices produced by ASR recognizers (Matusov et al., 2006; Mathias and Byrne, 2006; Zhou et al., 2007): most general & challenging
  - Confusion networks (Bertoldi et al., 2008): special case of lattice
    - Better trade-off to address the reordering issue
- SCFG models can take ASR lattice for ST
  - Generalized CKY algorithm for translating lattices (Dyer et al., 2008).
  - Nodes in the lattice numbered such that the end node is always numbered higher than the start node for any edge,
  - The vertex in HG labeled with $[X, i, j]$; here $i < j$ are the node numbers in the input lattice that are spanned by X.
- If reordering is critical in joint ST, try SCFG-based SMT models
  - Reordering without length constraints
  - Avoids traversing the lattice to compute the distortion costs

Coupling ASR and SMT for ST – Joint Modeling

1. Overview: ASR, SMT, ST, Metric
2. Learning Problems in SMT
3. Translation Structures for ST
4. Decoding: A Unified Perspective ASR/SMT
5. **Coupling ASR/SMT:** Decoding & **Modeling**
6. Practices
7. Future Directions

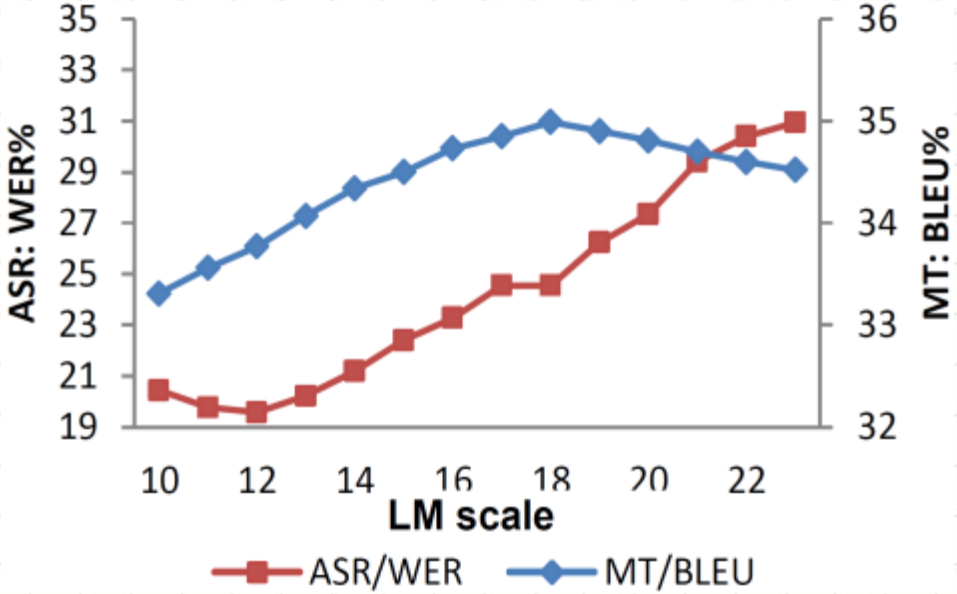
End-to-end Modeling of Speech Translation

- Two modules in conventional speech translation



- Problems of inconsistency
  - SR and MT are optimized for different criteria, inconsistent to the E2E ST quality (metric discrepancy)
  - SR and MT are trained without considering the interaction between them (train/test condition mismatch)

Why End-to-end Modeling Matters?



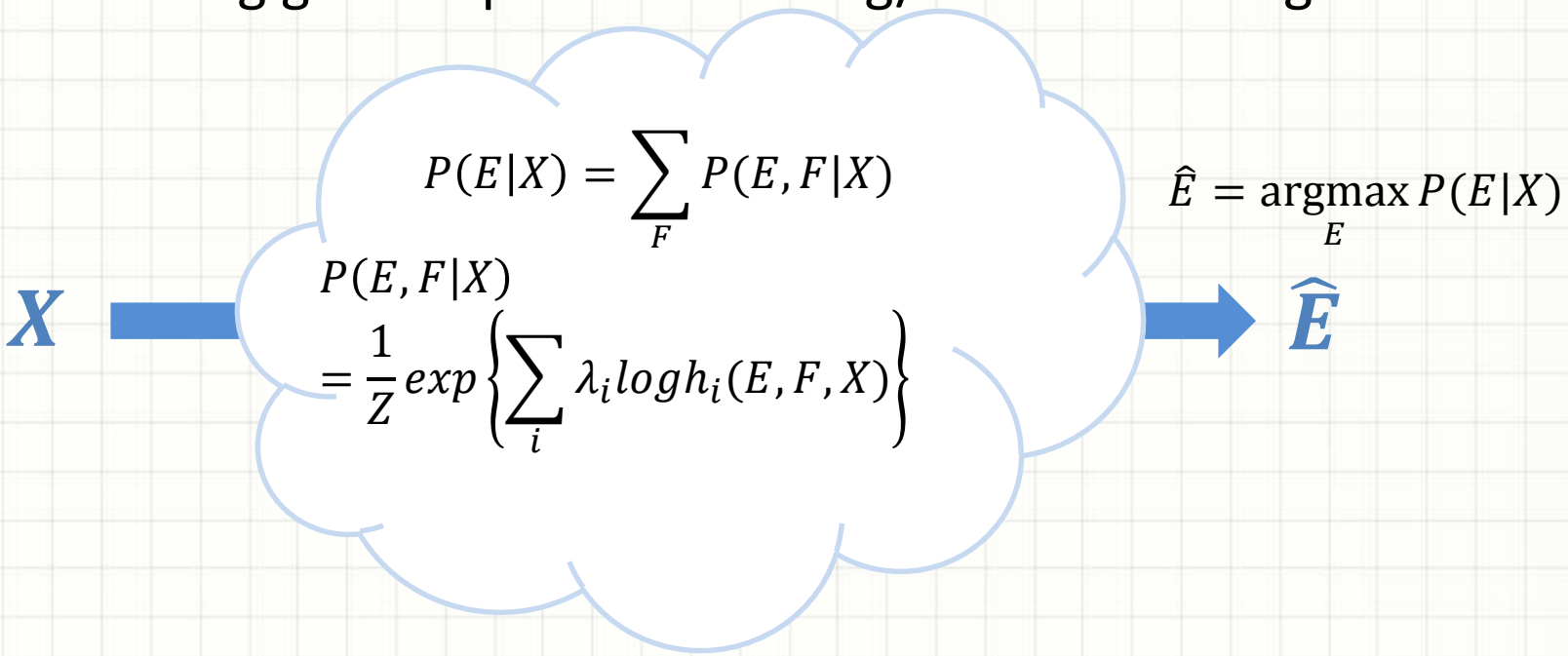
Lowest WER not necessarily gives the best BLEU

Fluent English is preferred as the input for MT, despite the fact that this might cause an increase of WER.

(He et al., 2011)

End-to-end ST Model

- A End-to-End log-linear model for ST
  - Enabling incorporation of rich features
  - Enabling a principal way to model the interaction between core components
  - Enabling global optimal modeling/feature training



Feature Functions for End-to-end ST

- Each feature function is a probabilistic model (except a few count features)
- Include all conventional ASR and MT models as features

features	model
Acoustic model	$h_{AM} = p(X\|F)$
Source language model	$h_{SLM} = P_{LM}(F)$
ASR hypothesis length	$h_{SWC} = e^{\|F\|}$
Forward phrase trans. Model	$h_{F2Eph} = \prod_k p(\tilde{e}_k\|\tilde{f}_k)$
Forward word trans. Model	$h_{F2Ewd} = \prod_k \prod_m \sum_n p(e_{k,m}\|f_{k,n})$
Backward phrase trans. Model	$h_{E2Fph} = \prod_k p(\tilde{f}_k\|\tilde{e}_k)$
Backward word trans. Model	$h_{E2Fwd} = \prod_k \prod_n \sum_m p(f_{k,n}\|e_{k,m})$
Translation Phrase count	$h_{PC} = e^{K}$
Translation hypothesis length	$h_{TWC} = e^{\|E\|}$
Phrase segment/reorder model	$h_{reorder} = P_{hr}(S\|E,F)$
Target language model	$h_{TLM} = P_{LM}(E)$

Learning Feature Weights In The Log-linear Model

- Jointly optimize the weights of features by MERT:

system	BLEU
Baseline (the current ASR and MT system)	33.8%
Global max-BLEU training (for all features)	35.2% (+1.4%)
Global max-BLEU training (for ASR-only features)	34.8% (+1.0%)
Global max-BLEU training (for SMT-only features)	34.2% (+0.4%)

Evaluated on a MS commercial data set

Case Study:

Transcript	it is great seeing you all here today
Translation ref.	今天很高兴在这里见到你们
Reco A.	let's great see you all here today
Translation A.	今天在这里看到你们让我们好
Reco B.	let's great to see you all here today
Translation B.	我们今天很高兴在这里见到你们

Reco. B contains one more *ins.* error. However, the insertion "to" :

i) makes the MT input grammatically more correct;

ii) provides critical context for "great";

iii) Provide critical syntactic info for word ordering of the translation.

Transcript	i didn't ever really wanna do this
Translation ref.	我从来没有真的想要这么做
Reco A.	i can never really wanna do this
Translation A.	我永远不能真的想
Reco B.	i ve never really want to do this
Translation B.	我从来没有真的要这么做

Reco. B contains two more errors. However, the mis-recognized phrase "want to" is plentifully represented in the formal text that is usually used for MT training and hence leads to correct translation.

Learning Parameters Inside The Features

- First, we define a generic differentiable utility function

$$U(\Lambda) = \sum_{r=1,\ldots R} \sum_{E_r \in hyp(F_r)} \sum_{F_r \in hyp(X_r)} p(E_r, F_r | X_r, \Lambda) \cdot C\left(E_r, E_r^*\right)$$

Λ : the set of model parameters that are of interest

X_r: the r-th speech input utterance

E_r^*: translation reference of the *r*-th utterance

E_r: translation hypothesis of the r-th utterance

F_r: speech recognition hypothesis of the r-th utterance

- $U(\Lambda)$ measures the end-to-end quality of ST, e.g.,
  - Choosing $C\left(E_r, E_r^*\right)$ properly, $U(\Lambda)$ covers a variety of ST metrics

$C\left(E_r, E_r^*\right)$	objective
$BLEU\left(E_r, E_r^*\right)$	Max expected BLEU
$1 - TER\left(E_r, E_r^*\right)$	Min expected Translation Error Rate

(He & Deng 2013)

Objective, Regularization, and Optimization

- Optimize the utility function directly
  - E.g., max expected BLEU
  - Generic gradient-based methods (Gao & He 2013)
  - Early-stopping/cross-validation on dev set
- Regularize the objective by K-L divergence
  - Suitable for parameters in a probabilistic domain
  - Training objective:

$$O(\Lambda) = logU(\Lambda) - \tau \cdot KL(\Lambda^0||\Lambda)$$

  - Extended Baum-Welch based optimization (He & Deng 2008, 2012, 2013)

EBW Formula for Translation Model

- Use lexicon translation model as an example

ASR score affects estimation of the translation model

$$p(g|h,\Lambda) = \frac{\sum_{k,m:\, f_{k,m}=g} \sum_{E,F} p(E,F|X,\Lambda')\Delta_E\, \gamma_h(k,m) + U(\Lambda')\tau_{FP} p(g|h,\Lambda^0) + D_h \cdot p(g|h,\Lambda')}{\sum_{k,m} \sum_{E,F} p(E,F|X,\Lambda')\Delta_E\, \gamma_h(k,m) + U(\Lambda')\tau_{FP} + D_h}$$

where $\Delta_E = [C(E) - U(\Lambda')]$, and $\gamma_h(k,m) = \frac{\sum_{n:e_{k,n}=h} p(f_{k,m}|e_{k,n},\Lambda')}{\sum_n p(f_{k,m}|e_{k,n},\Lambda')}$

Training is influenced by translation quality

Evaluation on IWSLT'11/TED

- Challenging open-domain SLT: TED Talks (www.ted.com)
  - Public speech on anything (tech, entertainment, arts, science, politics, economics, policy, environment …)
- Data (Chinese-English machine translation track)
  - **Parallel Zh-En** data: **110*K*** snt. **TED** transcripts; **7.7*M*** snt. **UN** corpus
  - **Monolingual English** data: **115*M*** snt. from Europarl, Gigaword, …
  - Example

English: What I'm going to show you first, as quickly as I can, is some foundational work, some new technology that we brought to ...
Chinese: 首先，我要用最快的速度为大家演示一些新技术的基础研究成果 …

Results on IWSLT'11/TED

- Phrase-based system
  - 1st phrase table from the TED parallel corpus
  - 2nd phrase table from 500K parallel snt selected from UN
  - 1st 3-gram LM from TED English transcription
  - 2nd 5-gram LM from 115M supplementary English snt
- Max-BLEU training only applied to the primary (TED) phrase table
  - Fine-tuning of full lambda set is performed at the end

BLEU scores on IWSLT test sets

system	Tst2010 (dev)	Tst2011 (test)
baseline	11.48%	14.68%
Max Expected BLEU training	12.39% (+0.9%)	15.92% (+1.2%)

← best single-system in IWSLT'11/CE_MT

(He & Deng 2012)

Further Reading

- F. Casacuberta, M. Federico, H. Ney and E. Vidal. 2008. Recent Efforts in Spoken Language Translation. IEEE Signal Processing Mag. May 2008
- C. Dyer, S. Muresan, and P. Resnik. 2008. Generalizing Word Lattice Translation. In Proc. ACL. 2008.
- L. Mathias and W. Byrne. 2006. Statistical phrase-based speech translation. In Proc. ICASSP. 2006. pp. 561–564.
- E. Matusov, S. Kanthak and H. Ney. 2006. Integrating speech recognition and machine translation: Where do we stand? In Proc. ICASSP. 2006.
- H. Ney. 1999. Speech translation: Coupling of recognition and translation. In Proc. ICASSP. 1999
- R. Zhang, G. Kikui, H. Yamamoto, T. Watanbe, F. Soong and W-K. Lo. 2004. A unified approach in speech-to-speech translation: Integrating features of speech recognition and machine translation. In Proc. COLING. 2004
- B. Zhou, L. Besacier and Y. Gao. 2007. On Efficient Coupling of ASR and SMT for Speech Translation. In Proc. ICASSP. 2007.
- B. Zhou, X. Cui, S. Huang, M. Cmejrek, W. Zhang, J. Xue, J. Cui, B. Xiang, G. Daggett, U. Chaudhari, S. Maskey and E. Marcheret. 2013. The IBM speech-to-speech translation system for smartphone: Improvements for resource-constrained tasks. Computer Speech & Language. Vol. 27, Issue 2, 2013, pp. 592–618

Further Reading

- X. He, A. Axelrod, L. Deng, A. Acero, M. Hwang, A. Nguyen, A. Wang, and X. Huang. 2011. The MSR system for IWSLT 2011 evaluation, in Proc. IWSLT, December 2011.
- X. He and L. Deng, 2013, Speech-Centric Information Processing: An Optimization-Oriented Approach, in Proceedings of the IEEE
- X. He, L. Deng, and A. Acero, 2011. Why Word Error Rate is not a Good Metric for Speech Recognizer Training for the Speech Translation Task?, in Proc. ICASSP, IEEE
- X. He, L. Deng, and W. Chou, 2008. Discriminative learning in sequential pattern recognition, IEEE Signal Processing Magazine, September 2008.
- Y. Liu , E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, M. Harper, 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies, IEEE Trans. on ASLP 2006.
- M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, J. Heng, J. Kahn, L. Yang, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wen, C. Woofers, 2008. Speech segmentation and spoken document processing, IEEE Signal Processing Magazine, May, 2008
- A. Waibel, C. Fugen, 2008. Spoken language translation, IEEE Signal Processing Magazine, vol.25, no.3, pp.70-79, May 2008.
- Y. Zhang, L. Deng, X. He, and A. Acero, 2011, A Novel Decision Function and the Associated Decision-Feedback Learning for Speech Translation, in ICASSP, IEEE

Practices

1. Overview: ASR, SMT, ST, Metric
2. Learning Problems in SMT
3. Translation Structures for ST
4. Decoding: A Unified Perspective ASR/SMT
5. Coupling ASR/SMT: Decoding & Modeling
6. **Practices**
7. Future Directions

Two techniques in depth:

1

- Domain adaptation

2

- System combination

Domain adaptation for ST

- Words differ in meaning across domains/contexts
- Domain adaptation is particularly important for ST
  - ST needs to handle spoken language
    - Colloquial style vs. written style
  - ST has interests on particular scenarios/domains
    - E.g., travel

Domain Adaptation by Data Selection for MT

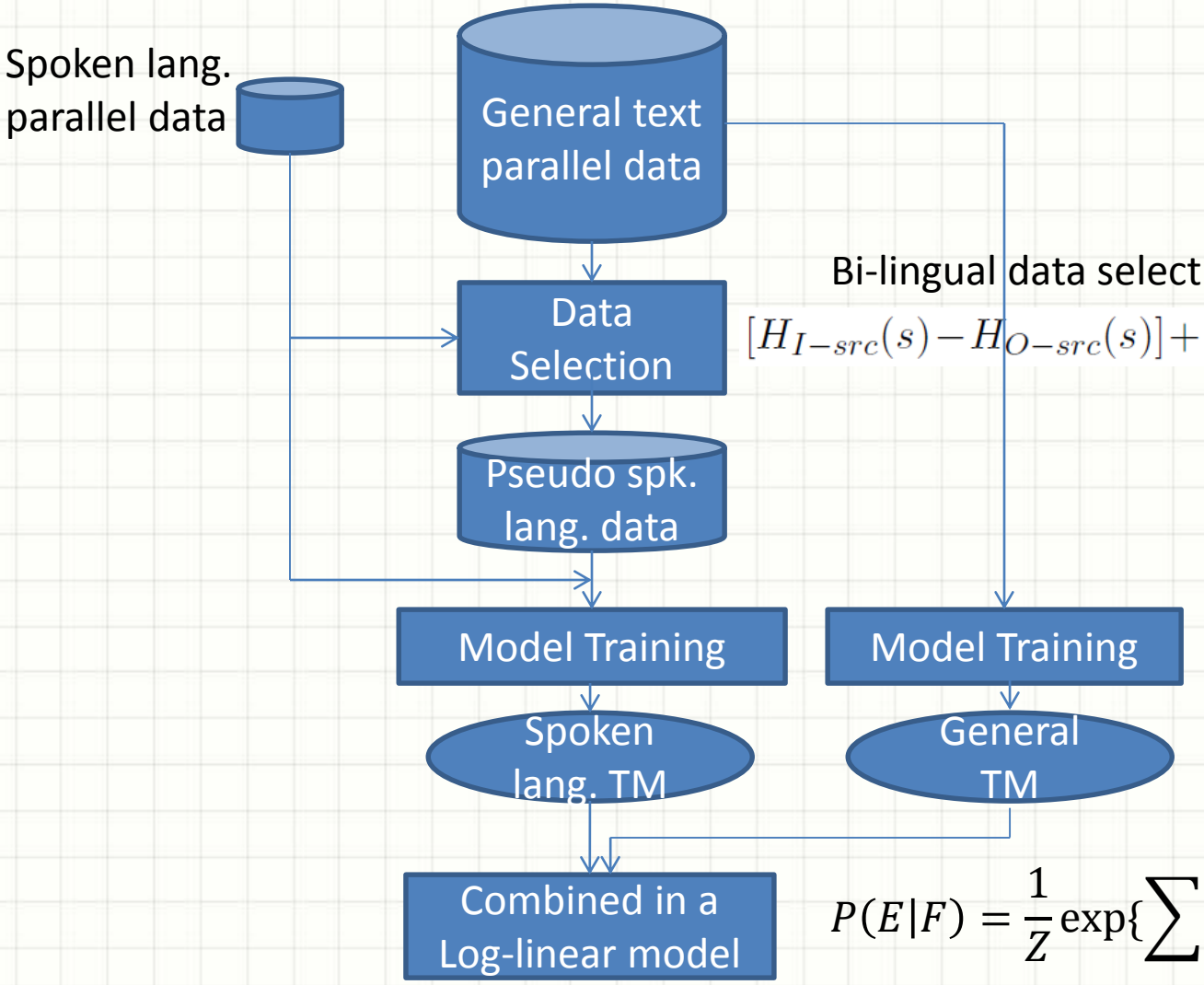
- Selecting data that match the targeting domain from a large general corpus
- MT needs data that match both *source* & *target* languages of the targeting domain.
  - The adapted system need to
    - cover domain-specific input
    - produce domain-appropriate output
  - e.g., "*Do you know of any restaurants* ***open*** *now ?*"
    - We need to select data to cover not only "open" at the source side, but also the right translation of "open" at the target side.
- Use bilingual cross-entropy difference:

$$\left[H_{in\_src}(s) - H_{out\_src}(s)\right] + \left[H_{in\_tgt}(s) - H_{out\_tgt}(s)\right]$$

The perplexity of *s* given the in-domain source language model "*in_src*"

(Axelrod, He, and Gao, 2011)

Multi-model and data selection based domain adaptation for ST



Bi-lingual data selection metric

$$[H_{I-src}(s) - H_{O-src}(s)] + [H_{I-tgt}(s) - H_{O-tgt}(s)]$$

$$P(E|F) = \frac{1}{Z}\exp\{\sum_i \lambda_i log f_i(E,F)\}$$

Evaluation on commercial ST data

- English-to-Chinese translation
  - Target: dialog style travel domain data
  - Training data available
    - 800K in-domain data
    - 12M general domain data
  - Evaluation
    - Performance on target domain
    - Performance on general domain, e.g., evaluate the robustness on out-of-target-domain.

Results and Analysis

Models and {λ} training	travel	general
General (dev: general)	16.22	18.85
Travel (dev: travel)	22.32 (+6.1)	10.81 (-8.0)
Multi-TMs (dev: travel)	22.12 (+5.9)	13.93 (-4.9)
Multi-TMs (dev: *tvl* : *gen* = *1:1*)	22.01 (+5.8)	16.89 (-2.0)
Multi-TMs (dev: *tvl* : *gen* = *1:2*)	20.24 (+4.0)	18.02 (-0.8)

Results reported in BLEU %

1) Using multiple translation models together as features helps
2) The feature weights, determined by the dev set, play a critical role
3) Big gain on in-domain test (travel), robust on out-of-domain test (general)

(From He & Deng, 2011)

Case Studies

Source English	General model (Chinese)	Travel-domain adapted model
A glass of cold water, please .	一杯冷水请。	请来一杯冷的水。
Please keep the **change** .	请保留更改。	不用找了。
Thank , this is for your **tip** .	谢谢，这是为您**提示**。	谢谢，这是给你的**小费**。
Do you know of any restaurants **open** now ?	你现在知道的任何**打开**的餐馆吗？	你知道现在还有餐馆在**营业**的吗？
I'd like a restaurant with cheerful atmosphere	我想就食肆的愉悦的气氛	我想要一家气氛活泼的餐厅。

Note the improved translation of ambiguous words (e.g., open, tip, change), and improved processing of colloquial style grammar (e.g., *a glass of cold water, please.*)

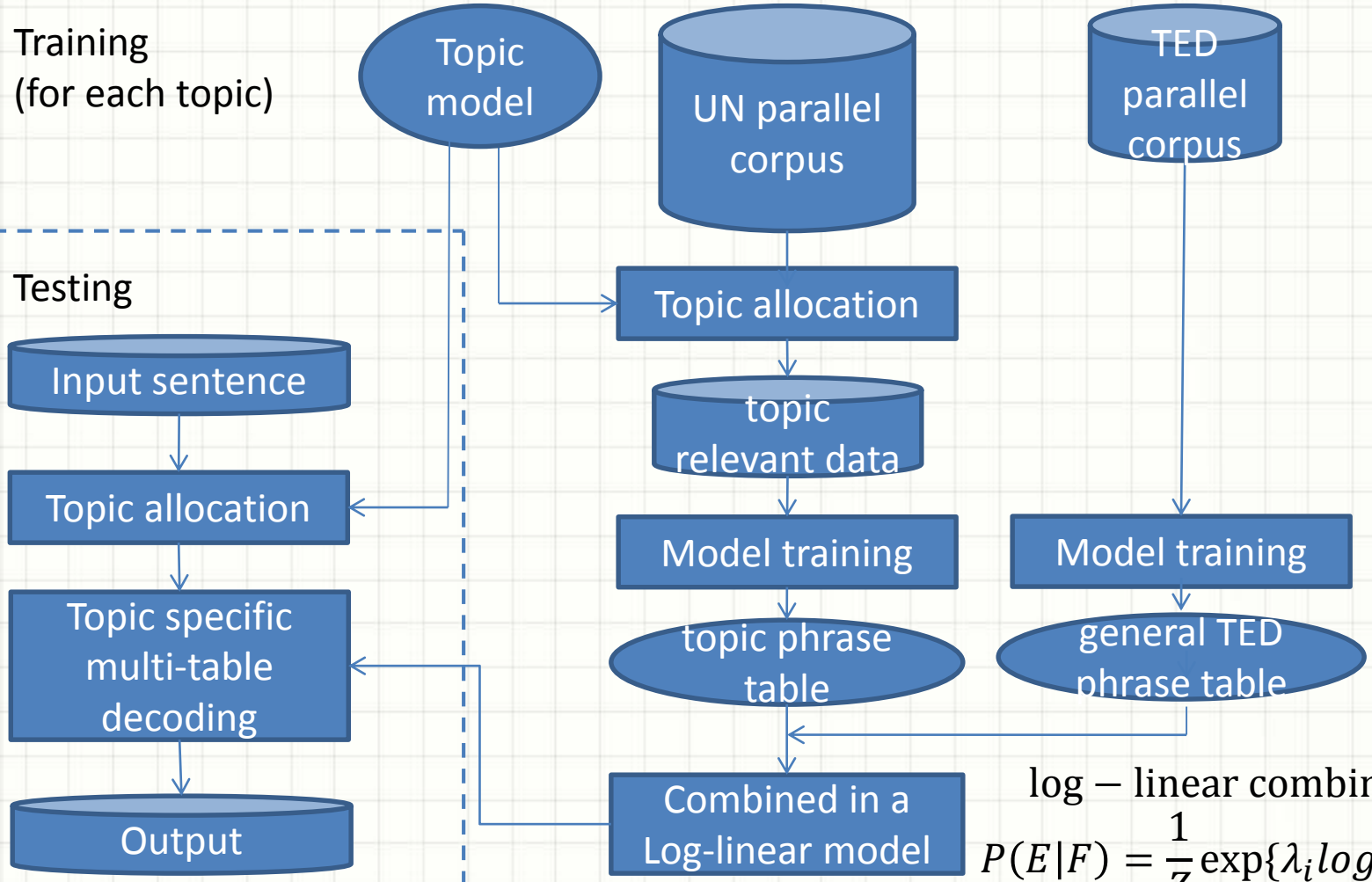
From domain adaptation to topic adaptation

- Motivation
  - Topic changes talk to talk, dialog to dialog
    - Meanings of words changes, too.
  - Lots of out-of-domain data
    - Broad coverage, but not all of them are relevant.
    - How to utilize the OOD corpus to enhance the translation performance?
- Method
  - Build topic model on target domain data
  - Select topic relevant data from OOD corpus
  - Combine topic specific model with the general model at testing

Analysis on IWSLT'11: topics of TED talks

- Build LDA-based topic model on TED data
  - estimated on the source side of the parallel data
- Restricted to 4 topics – are they meaningful?
  - Only has 775 talks/110K sentences
- Look at some of the top keywords in each topic:
  - Topic 1 (design, computer, data, system, machine) ***technology***
  - Topic 2 (Africa, dollars, business, market, food, China, society) ***global***
  - Topic 3 (water, earth, universe, ocean, trees, carbon, environment) ***planet***
  - Topic 4 (life, love, god, stories, children, music) ***abstract***
- Reasonable clustering

Topic modeling, data selection, and multi-phrase-table decoding



Experiments on IWSLT'11/TED

- For each topic , select 250~400K sentences from the UN corpus, train a topic-specific phrase table
- Evaluation results on IWSLT'11 (TED talk dataset)
  - Simply adding an extra UN-driven phrase table didn't help
  - Topic specific multi-phrase table decoding helps

Phrase table used	Dev (2010)	Test (2011)
TED only (baseline)	11.3%	13.0%
TED + UN-all	11.3%	13.0%
TED + UN-4 topics	11.8%	13.5%

(From Axelrod et al., 2012)

System Combination for MT

Hypotheses from single systems

E_1: she bought the Jeep

E_2: she buys the SUV

E_3: she bought the SUV Jeep

Combined MT output

she bought the Jeep SUV

- System Combination:
  - *Input*:

    a set of translation hypotheses from multiple single MT systems, for the same source sentence
  - *Output*:

    a *better* translation output derived from input hypotheses

Confusion Network Based Methods

Hypotheses from single systems

E_1: she bought the Jeep

E_2: she buys the SUV

E_3: she bought the SUV Jeep

Combined MT output

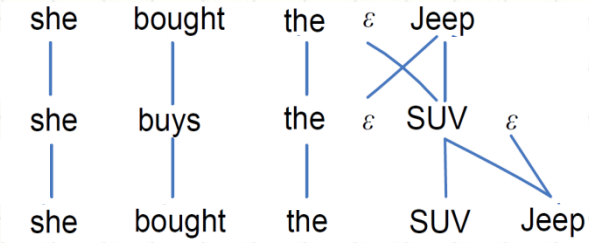
she bought the SUV

1) *Select the backbone*

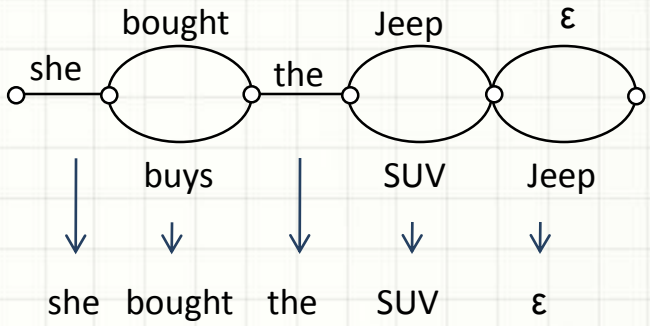
$$E_B = \arg\min_{E' \in \mathbf{E}_B} \sum_{E \in \mathbf{E}_e} P(E \mid F) L(E', E)$$

e.g., E_2 is selected

2) *Align hypotheses*



3) *Construct and decode CN*



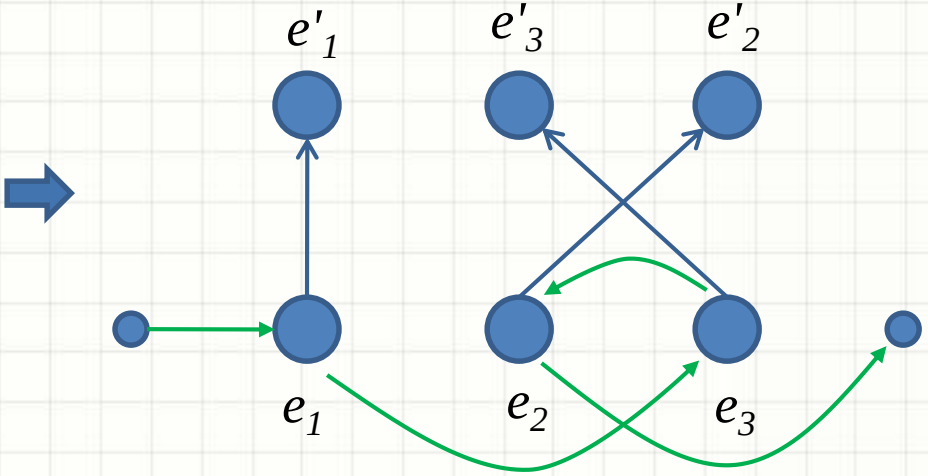
Hypothesis Alignment

- Hypothesis alignment is crucial
  - GIZA based approach (Matusov et al., 2006)
    - Use GIZA as a generic tool to align hypotheses
  - TER based alignment (Sim et al., 2007, Rosti et al. 2007)
    - Align one hypothesis to another such that the TER is minimized
  - HMM based hypothesis alignment (He et al. 2008, 2009)
    - Use fine-grained statistical model
    - No training needed, HMM's parameters are derived from pre-trained bilingual word alignment models
  - ITG based approach (Karakos et al., 2008)
  - A latest survey and evaluation (Rosti et al., 2012)

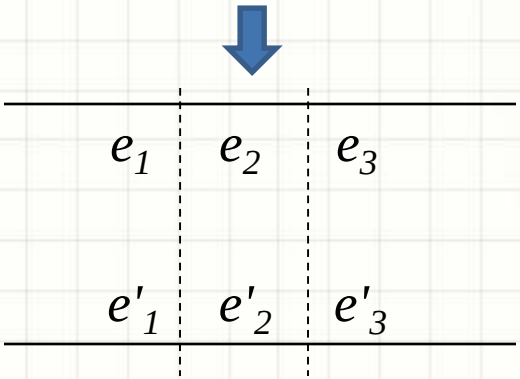
HMM based Hypothesis Alignment

E_B : e_1 e_2 e_3

E_{hyp}: e'_1 e'_3 e'_2



- HMM is built on the backbone side
- HMM aligns the hypothesis to the backbone
- After alignment, a CN is built

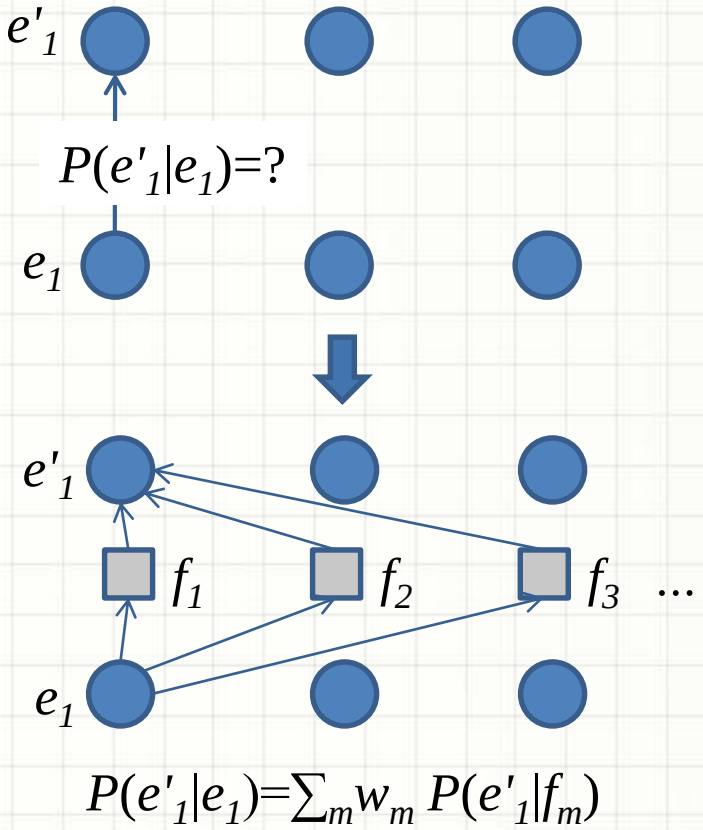


HMM Parameter Estimation

- Emitting Probability (via words in source sentence)
  - $P(e'_1|e_1)$ models how likely e'_1 and e_1 have similar meanings
  - Use the source word sequence $\{f_1, \ldots, f_M\}$ as a hidden layer, $P(e'_1|e_1)$ takes a mixture-model form, i.e.,

    $$P_{src}(e'_1|e_1)=\sum_m w_m P(e'_1|f_m)$$

    where $w_m = P(f_m|e_1)/\sum_m P(f_m|e_1)$
  - $P(e'_1|f_m)$ is from the bilingual word alignment model, F→E direction
  - $P(f_m|e_1)$ is from that of E→F

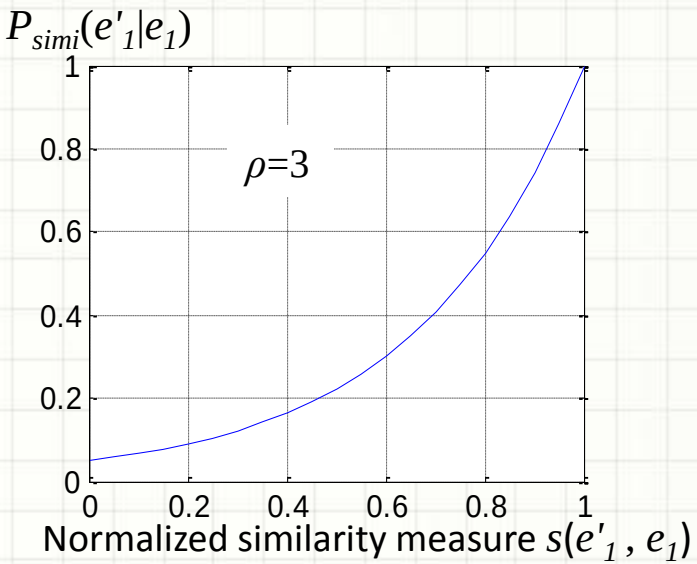


HMM Parameter Estimation (cont.)

- Emitting Probability (via word surface similarity)
  - Normalized similarity measure s
    - Based on Levenshtein distance
    - Based on matched prefix length
  - Use an exponential mapping to get $P(e'_1|e_1)$

$$P_{simi}(e'_1|e_1)=\exp[\rho \cdot(s\,(e'_1, e_1) -1)]$$

$s\,(e'_1, e_1)$ is normalized to [0,1]



- Overall Emitting Probability

$$P(e'_1|e_1) = \alpha \cdot P_{src}(e'_1|e_1)+(1- \alpha)\cdot P_{simi}(e'_1|e_1)$$

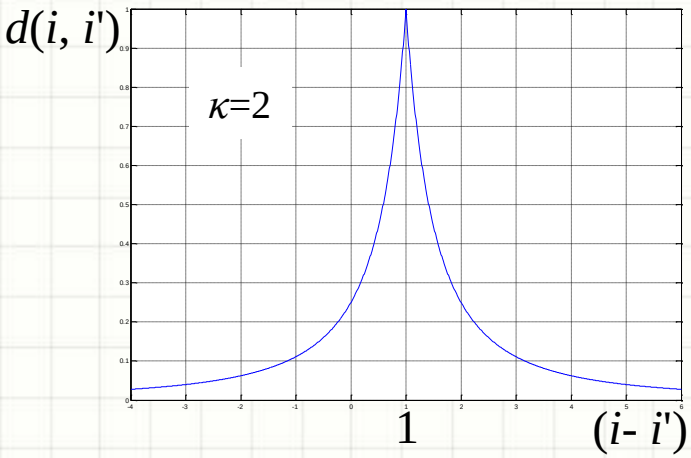
HMM Parameter Estimation (cont.)

- Transition Probability
  - $P(a_j|a_{j-1})$ models word ordering
    - Takes the same form as a bilingual word alignment HMM
    - Strongly encourages monotonic word ordering
    - Allows non-monotonic word ordering

$$d(i,i') = \left(1 + |i - i' - 1|\right)^{-\kappa}$$

$$P(a_j \mid a_{j-1}) = \frac{d(a_j, a_{j-1})}{\sum_i d(i, a_{j-1})}$$

a_j – alignment of the j-th word



Find the Optimal Alignment

- Viterbi decoding:

$$\hat{a}_1^J = \arg\max_{a_1^J} \prod_{j=1}^{J} \left[p(a_j \mid a_{j-1}, I)\, p(e'_j \mid e_{a_j}) \right]$$

- Other variations:
  - posterior probability & threshold based decoding
  - max posterior mode decoding

Decode the Confusion Network

- Log-linear model based decoding (Rosti et al. 2007)
  - Incorporate multiple features (e.g., voting, LM, length, etc.)

$$E^* = \arg\max_{E' \in \mathbf{E}_h} \ln P(E' \mid F)$$

where

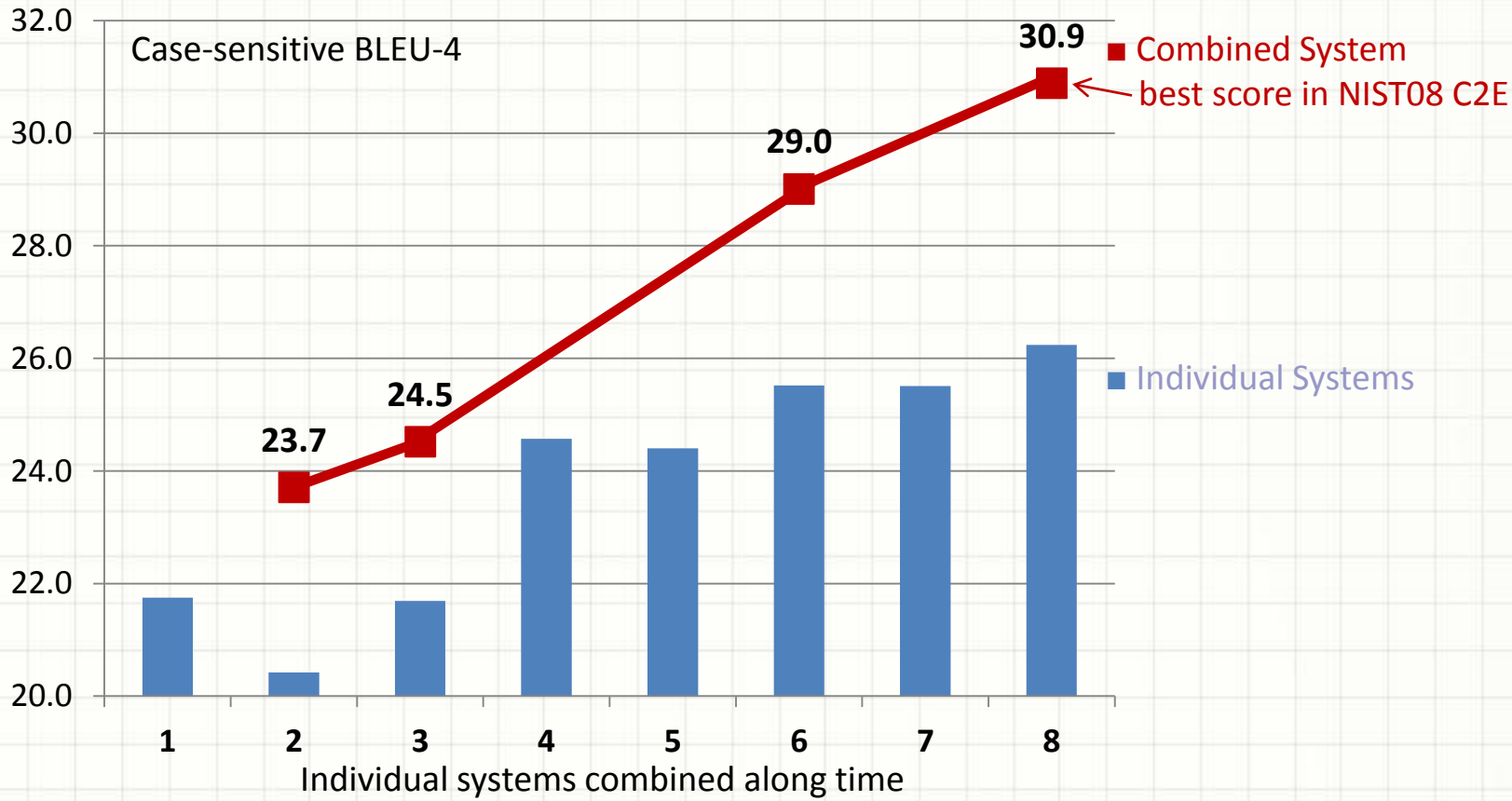
$$\ln P(E' \mid F) = \ln \prod_{l=1}^{L} P_{S-MBR}(e'_l \mid F) + \nu \ln P_{LM}(E') + \xi \mid E' \mid$$

Confusion network decoding (L=5)

he	have	ε	good	car
he	has	ε	nice	sedan
it	ε	a	nice	car
he	has	a	ε	sedan
e_1	e_2	e_3	e_4	e_5

Results on 2008 NIST Open MT Eval

The MSR-NRC-SRI entry for Chinese-to-English



(from He et al., 2008)

Related Approaches & Extensions

- Incremental hypothesis alignment
  - (Rosti et al., 2008; Li et al, 2009, Karakos et al., 2010)
- Other approaches
  - Joint decoding (He and Toutanova, 2009)
  - Phrase-level system combination (Feng et al, 2009)
  - System combination as target-to-target decoding (Ma & McKeown, 2012)
- Survey and evaluation
  - (Rosti et al., 2012)

Comparison of aligners

Results reported in BLEU % (best scores in bold)
Significance group ID is marked as superscripts on results

Aligner	NIST MT09 Arabic-English	WMT11 German-English	WMT11 Spanish-English
Best single system	51.74	24.16	30.14
GIZA	57.95[1]	26.02[1]	33.62[1]
Incr. TER	58.63[2]	**26.39[2]**	33.79[1]
Incr. TER++	59.05[2]	26.10[1]	33.61[1]
Incr. ITG++	**59.37[3]**	**26.50[2]**	33.85[1]
Incr. IHMM	**59.27[3]**	**26.40[2]**	**34.05[2]**

(From Rosti, He, Karakos, Leusch, et al., 2012)

Error Analysis

- Per-sentence errors were analyzed
  - Paired Wilcoxon test measured the probability that the error reduction relative to the best system was real
- Observations
  - All aligners significantly reduce substitution/shift errors
  - No clear trend for insertions/deletions – sometimes worse than the best system
- Further analysis showed that agreement for the non-NULL words is important measure for alignment quality, but agreement on NULL tokens is not

Further Reading

- Y. Feng, Y. Liu, H. Mi, Q. Liu and Y. Lü , 2009, Lattice-based system combination for statistical machine translation, in Proceedings of EMNLP
- X. He and K. Toutanova, 2009, Joint Optimization for Machine Translation System Combination, In Proceedings of EMNLP
- X. He, M. Yang, J. Gao, P. Nguyen, and R. Moore, 2008, Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems, in Proceedings of EMNLP
- D. Karakos, J. Eisner, S. Khudanpur, and M. Dreyer. 2008. Machine Translation System Combination using ITG-based Alignments. In Proceedings of ACL-HLT.
- D. Karakos, J. Smith, and S. Khudanpur. 2010. Hypothesis ranking and two-pass approaches for machine translation system combination. In Proc. ICASSP.
- C. Li, X. He, Y. Liu, and N. Xi, 2009, Incremental HMM Alignment for MT System Combination, in Proceedings of ACL-IJCNLP
- W. Ma and K McKeown. 2012. Phrase-level System Combination for Machine Translation Based on Target-to-Target Decoding. In Proceedings of AMTA
- E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In Proceedings of EACL.
- A. Rosti, B. Xiang, S. Matsoukas, R. Schwartz, N. Fazil Ayan, and B. Dorr. 2007. Combining outputs from multiple machine translation systems. In Proceedings of NAACL-HLT.
- A. Rosti, B. Zhang, S. Matsoukas, and R. Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In Proceedings of WMT
- A. Rosti, X. He, D. Karakos, G. Leusch, Y. Cao, M. Freitag, S. Matsoukas, H. Ney, J. Smith, and B. Zhang, Review of Hypothesis Alignment Algorithms for MT System Combination via Confusion Network Decoding , in Proceedings of WMT
- K. Sim, W. Byrne, M. Gales, H. Sahbi, and P. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In Proceedings of ICASSP.
- A. Axelrod, X. He, and J. Gao, Domain Adaptation via Pseudo In-Domain Data Selection, in Proc. EMNLP, 2011
- A. Bisazza, N. Ruiz and M. Federico , 2011, Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation, in Proceedings of IWSLT
- G. Foster and R. Kuhn. 2007. Mixture-Model Adaptation for SMT. Workshop on Statistical Machine Translation, in Proceedings of ACL
- G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In Proceedings of EMNLP.
- B. Haddow and P. Koehn, 2012. Analysing the effect of out-of-domain data on smt systems. in Proceedings of WMT
- X. He and L. Deng, 2011, Robust Speech Translation by Domain Adaptation, in Proceedings of Interspeech.
- P. Koehn and J. Schroeder, 2007. Experiments in domain adaptation for statistical machine translation, in Proceedings of WMT.
- S. Mansour, J. Wuebker and H. Ney , 2012, Combining Translation and Language Model Scoring for Domain-Specific Data Filtering, in Proceedings of IWSLT
- R. Moore and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. Association for Computational Linguistics.
- P. Nakov, 2008. Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing, in Proceedings of WMT

Summary

- Fundamental concepts and technologies in speech translation
- Theory of speech translation from a joint modeling perspective
- In-depth discussions on domain adaptation and system combination
- Practical evaluations on major ST/MT benchmarks

ST Future Directions --Additional to ASR and SMT

- Overcome the low-resource problem:
  - How to *rapidly* develop ST with limited amount of data, and adapt SMT for ST?
- Confidence measurement:
  - Reduce misleading dialogue turns & lower the risk of miscommunication.
  - Complicated by the combination of two sources of uncertainty in ASR and SMT.
- Active vs. Passive ST
  - More active, e.g., to actively clarify, or warn the user when confidence is low
  - Development of suitable dialogue strategies for ST.
- Context-aware ST
  - ST on smart phones (e.g., location, conversation history).
- End-to-end modeling of speech translation with new features
  - E.g., prosody acoustic features for translation via end-to-end modeling
- Other applications of speech translation
  - E.g., cross-lingual spoken language understanding

Further Reading About this Tutorial

- This tutorial is mainly based on:
  - Bowen Zhou, Statistical Machine Translation for Speech: A Perspective on Structure, Learning and Decoding, in *Proceedings of the IEEE*, IEEE, May 2013
  - Xiaodong He and Li Deng, Speech-Centric Information Processing: An Optimization-Oriented Approach, in *Proceedings of the IEEE*, IEEE, May 2013
- Both provide references for further reading
- Papers and charts available on authors' webpages

Resources (play it yourself)

- You need also an ASR system
  - HTK: http://htk.eng.cam.ac.uk/
  - Kaldi: http://kaldi.sourceforge.net/
  - HTS (TTS for S2S): http://hts.sp.nitech.ac.jp/
- MT Toolkits (word alignment, decoder, MERT)
  - GIZA++: http://www.statmt.org/moses/giza/GIZA++.html
  - Moses: http://www.statmt.org/moses/
  - Cdec: http://cdec-decoder.org/
- Benchmark data sets
  - IWSLT: http://www.iwslt2013.org
  - WMT: http://www.statmt.org (include Europarl)
  - NIST Open MT: http://www.itl.nist.gov/iad/mig//tests/mt/

Q & A

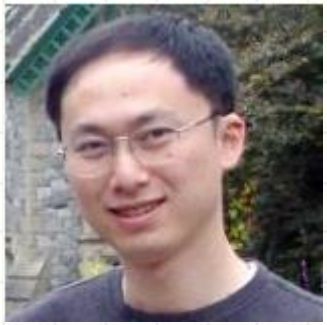
Bowen Zhou

RESEARCH MANAGER

IBM WATSON RESEARCH CENTER

http://researcher.watson.ibm.com/researcher/view.php?person=us-zhou

LinkedIn http://www.linkedin.com/pub/bowen-zhou/18/b5a/436



Xiaodong He

RESEARCHER

MICROSOFT RESEARCH REDMOND

AFFILIATE PROFESSOR

UNIVERSITY OF WASHINGTON, SEATTLE

http://research.microsoft.com/~xiaohe