Integrating Automatic Speech Recognition and Statistical Machine Translation

Bill Byrne

Cambridge University Engineering Department

Center for Language and Speech Processing
The Johns Hopkins University

1 April 2006

Work done with Shankar Kumar, now at Google, Inc.
Yonggang Deng, now at IBM T.J. Watson
Lambert Mathias, JHU CLSP

Outline

- Introduction
- TTM – Phrase Translation via WFSTs
  - WFST Operations
  - TTM
  - Phrase Pairs
  - Phrase Segmentation
  - Phrase Translation
  - Phrase Insertion
  - Phrase Reordering
- TTM – Speech Translation
  - A Generative Model of Speech Translation
  - Translation of ASR Lattices
- Alignment Models
  - Model-4
  - Word-to-Word HMM
  - Word-to-Phrase Alignment, Phrase Pair Posteriors, Bigram Translation
  - Translation Performance
- Discussion
  - Announcements
  - Unsolicited Opinions

MT Viewed From an ASR Perspective – Two Topics

Integrating Automatic Speech Recognition and Statistical Machine Translation

(1) Speech Translation

(2) Application of ASR modeling approaches to MT

- Word Alignment
- Translation

Can We Pretend that MT is ASR ?

Of course. We just need :

- Training data: parallel text, monolingual text
- Alignment models and estimation algorithms
- Search algorithms for translation
- Some way to measure translation quality

Five Easy Problems in Statistical Machine Translation

What's currently needed to build a basic phrase-based SMT system:

1. Parallel Text for SMT Training: Document and Sentence Alignment
2. Accurate and Computationally Efficient Models of Word and Phrase Alignment
3. Language Models, Estimated from large amounts of monolingual text
4. Translation – Model-Based Search Algorithms
5. Automatic Measurements of Translation and Alignment Quality
    - Support iterative system development
    - Useful for task-specific translation strategies, i.e. hypothesis and system combination

Translation – Lessons from ASR System Development Focus on Modeling and Avoid Specialized Search Algorithms

Once the models are defined, translation is 'trivial' . Its just search ...

Translate a target language sentence **T** into a source language sentence **S** :

$$\widehat{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmax}}\ P(\mathbf{S}\,|\,\mathbf{T}) = \underset{\mathbf{S}}{\operatorname{argmax}}\ \underbrace{P(\mathbf{T}\,|\,\mathbf{S})}_{\text{Translation Model}}\ \underbrace{P(\mathbf{S})}_{\text{Language Model}}$$

One approach is to construct specialized search algorithms

- Depending on the underlying models, search can be by Viterbi, A^*, or other specialized search procedures

But decoder design and implementation is complex

- Small model changes might require large changes to a decoder
- Approximations in search imply inexact implementation of the models
  - Can only implement what can be implemented
  - May as develop models which lead to exact realizations
- Decoder implementation takes effort away from 'modeling'

Translation from Speech is Also Trivial

A model-based approach to translation is easy to formulate

- Foreign Language Speech - A

- ASR Acoustic Model - $P(A \mid \mathbf{T})$

- Provided by a foreign language ASR system, e.g. using acoustic HMMs

Transcription Followed by Translation

$$A \underset{\substack{\text{ASR}\\ \text{System}}}{\Longrightarrow} \widehat{\mathbf{T}}$$

$$\widehat{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmax}}\, P(\,\widehat{\mathbf{T}} \mid \mathbf{S}\,)\, P(\mathbf{S})$$

Integrated Speech Translation

$$\widehat{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmax}} \sum_{\mathbf{T}} P(A \mid \mathbf{T})\, P(\mathbf{T} \mid \mathbf{S})\, P(\mathbf{S})$$

If we have the ASR and the translation models, this is ‘merely’ a search problem ...

Outline

Introduction

TTM – Phrase Translation via WFSTs

- WFST Operations
- TTM
- Phrase Pairs
- Phrase Segmentation
- Phrase Translation
- Phrase Insertion
- Phrase Reordering

TTM – Speech Translation

- A Generative Model of Speech Translation
- Translation of ASR Lattices

Alignment Models

- Model-4
- Word-to-Word HMM
- Word-to-Phrase Alignment, Phrase Pair Posteriors, Bigram Translation
- Translation Performance

Discussion

- Announcements
- Unsolicited Opinions

Translation via Weighted Finite State Transducers

Translation with Finite State Devices [1]

- Implements the simpler IBM word alignment models as WFSTs
  - word-to-word translation, word fertility, and permutation (reordering)

If the component models can be implemented as WFSTs which can be composed, building a decoder is trivial

- The value of this modeling approach has been shown in ASR by the systems developed at AT&T [2]
- Translation is performed using libraries of standard FSM operations
- Architectures can be limited, but avoids special-purpose decoders
- Clear formulation of underlying model components
- Easy to work on translation components in isolation

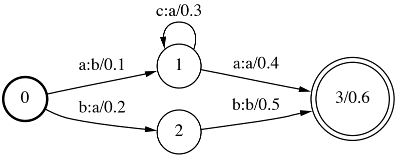
There are a variety of FSM-based approaches to translation [3]

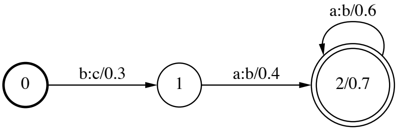
[1] K. Knight and Y. Al-Onaizan (1998), Translation with Finite-State Devices, Proc. AMTA.

[2] M. Mohri and M. Riley (1999), Integrated context-dependent networks in very large vocabulary speech recognition, EUROSPEECH.

[3] S. Bangalore and G. Riccardi (2001), A finite-state approach to machine translation. Proc. NAACL.

Weighted Finite State Transducer Operations [4]





Transducer A

b b → a b / 1.3
a a → b a / 1.1
a c a → a a a / 1.4
a c c a → a a a a / 1.7
...

Transducer B

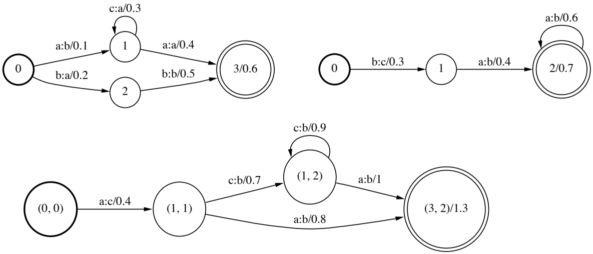
b a → c b / 1.4
b a a → c b b / 2.0
b a a a → c b b b / 2.6
...

Transducers can realize Markov processes : $-\log P_A(\text{ b a } | \text{ a a }) = 1.1$
$-\log P_B(\text{ c b } | \text{ b a }) = 1.4$
...

The maximum likelihood mapping is the path with the shortest cost

[4] M. Mohri and M. Riley (2002), Weighted Finite-State Transducers in Speech Recognition (Tutorial). International Conference on Spoken Language Processing. http://www.research.att.com/projects/mohri/postscript/icslp.ps

Weighted Finite State Transducer Operations



- Transducer C is the composition of transducers A and B : $C = A \cdot B$
- C maps a string x to a string y through an intermediate, hidden string z

$$x \longrightarrow_A \longrightarrow z \longrightarrow_B \longrightarrow y$$

$$\max_y P_C(y|x) = \max_y \max_z P_B(y|z) \, P_A(z|x)$$

- Natural relationship between Markov processes and transducers

TTM – Transducer Translation Model

Generative source-channel model of machine translation

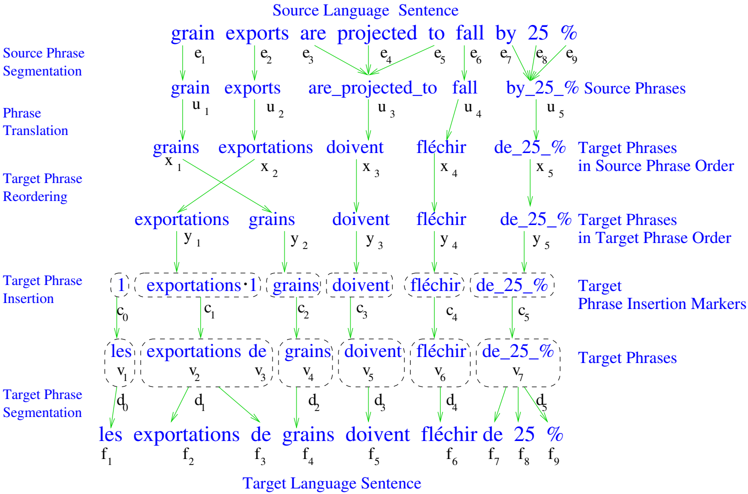
- Takes the best of
  - Och & Ney's Phrase-based translation models [5]
  - Knight & Al Onaizan WFST description of translation via IBM models
- Word alignment and translation under the model can be performed using standard WFST operations
  - Modular implementation
  - No need for a specialized decoder - "the model is the decoder"
  - Can easily generate translation lattices and N-best lists

Overall Goals :

- Relatively good performance with models that are really quite simple
- Should be easy to apply to translation of ASR lattices
- Easy to learn, easy to modify

[5] F. Och (2002), Statistical Machine Translation: From Single Word Models to Alignment Templates. PhD. Thesis, RWTH Aachen, Germany

TTM – Translation with Moving Target Phrase Order

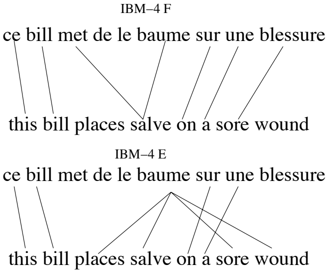

Source Language Sentence
grain exports are projected to fall by 25 %
e_1 e_2 e_3 e_4 e_5 e_6 e_7 e_8 e_9
Source Phrase Segmentation
grain exports are_projected_to fall by_25_% Source Phrases
u_1 u_2 u_3 u_4 u_5
Phrase Translation
grains exportations doivent fléchir de_25_% Target Phrases in Source Phrase Order
x_1 x_2 x_3 x_4 x_5
Target Phrase Reordering
exportations grains doivent fléchir de_25_% Target Phrases in Target Phrase Order
y_1 y_2 y_3 y_4 y_5
Target Phrase Insertion
1 exportations·1 grains doivent fléchir de_25_% Target Phrase Insertion Markers
c_0 c_1 c_2 c_3 c_4 c_5
les exportations de grains doivent fléchir de_25_% Target Phrases
v_1 v_2 v_3 v_4 v_5 v_6 v_7
Target Phrase Segmentation
d_0 d_1 d_2 d_3 d_4 d_5
les exportations de grains doivent fléchir de 25 %
f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9
Target Language Sentence

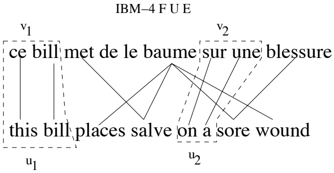
- Transformations via stochastic models implemented as WFSTs
- Built with standard WFST operations such as composition and best-path search

Phrase Pairs from Word Aligned Parallel Text

Alignment Template Models (Och et al. 1999)

- Derived from ‘good’ word-level alignments, typically from IBM-4
  - more recently by Word-to-Phrase Alignment HMMs
- Alignments and models are produced in both translation directions


IBM–4 F
ce bill met de le baume sur une blessure
this bill places salve on a sore wound
IBM–4 E
ce bill met de le baume sur une blessure
this bill places salve on a sore wound


IBM–4 F U E
v1
v2
ce bill met de le baume sur une blessure
this bill places salve on a sore wound
u1
u2

- Phrase Pairs are extracted to cover word alignment patterns
  - Phrase – a word sequence that can be translated
- Probability distributions are defined over phrase pair sequences
- Primarily interested in covering the phrases in the test set

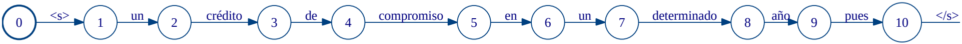
The Phrase Pair Inventory

Phrase pairs are extracted from word-aligned parallel text

English Phrase u	French Phrase v	Training Instances	Phrase Transduction Probability $P(v\|u)$
hear_hear	bravo	16	0.8
	bravo_bravo	3	0.15
	ordre	1	0.05
terms_of_reference	mandat	4	0.8
	de_son_mandat	1	0.2

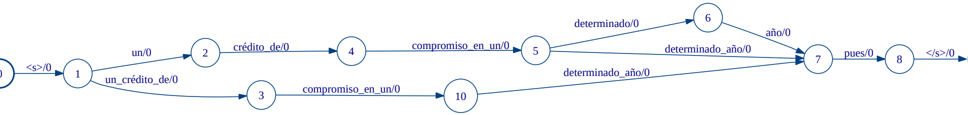
- A good Phrase Pair Inventory is crucial for translation
- Two important considerations :
  - Word Alignment Quality of underlying models → yields good phrase pairs
  - Coverage of phrases in the test set
    - test set phrases can be translated only if they can be found in the word-aligned parallel text

Target Phrase Segmentation



Sentence Acceptor

⇓ Target Phrase Segmentation Transducer

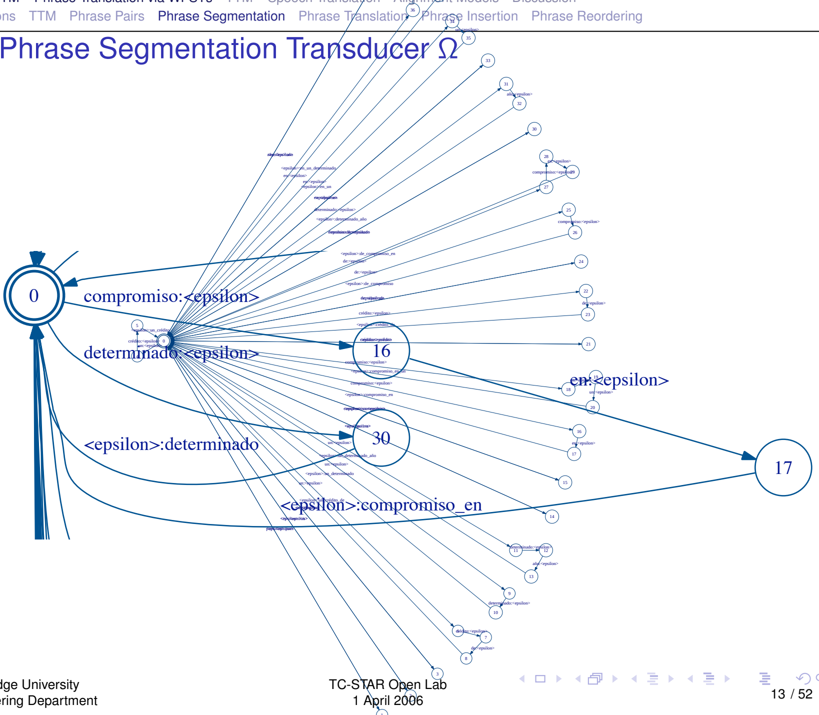


Phrase Sequence Lattice Ω

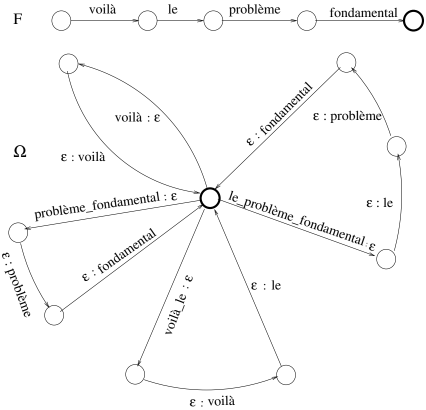
Phrase Sequence Lattice contains the phrase sequences in the foreign text

- all phrase sequences correspond to the single foreign sentence
- recall, a phrase is a sequence of words which can be translated
- different phrase sequences lead to different translations
- the lattice is unweighted

Target Phrase Segmentation Transducer Ω



Target Phrase Segmentation Transducer Ω



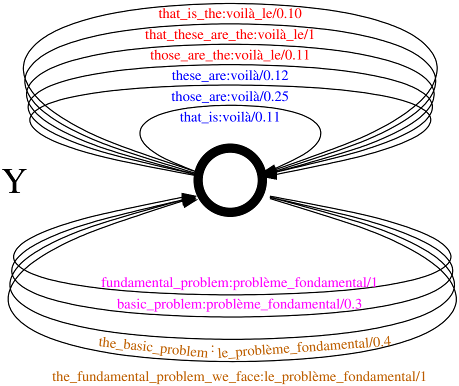
In translation of text, this transducer implements a degenerate distribution:

$$P(T|v_1^K) = \begin{cases} 1 & T \sim v_1^K \\ 0 & \text{other} \end{cases}$$

where v_1^K is any phrase sequence

Phrase Translation Transducer *Y*

Single state, trivial transducer to implement phrase sequence translation


Y
that_is_the:voilà_le/0.10
that_these_are_the:voilà_le/1
those_are_the:voilà_le/0.11
these_are:voilà/0.12
those_are:voilà/0.25
that_is:voilà/0.11
fundamental_problem:problème_fondamental/1
basic_problem:problème_fondamental/0.3
the_basic_problem : le_problème_fondamental/0.4
the_fundamental_problem_we_face:le_problème_fondamental/1

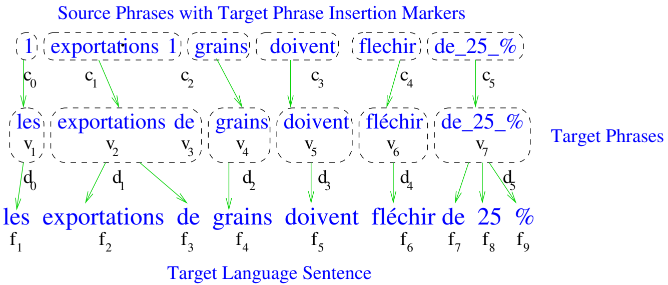
- Maps English phrases into French

- Based on the Phrase pair Inventory

- Phrase sequences are translated phrase–by–phrase

$$P(v_1^K | u_1^K) = \prod_{k=1}^{K} p(v_k | u_k)$$

Target Language Phrase Insertion



Different phrase segmentations lead to different translations :

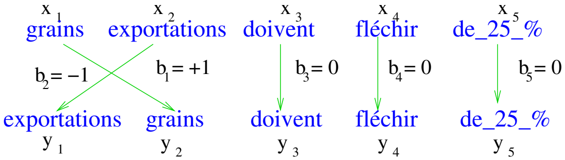
- Sequences c_0^K that could have generated F : $Y \circ \Omega \circ F$

H1 that_these_are_the fundamental_problem
voilà_le problème_fondamental

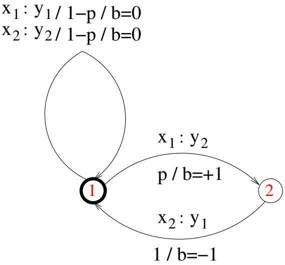
.

H16 that_is the_basic_problem
voilà le_problème_fondamental

Phrase Swapping by WFSTs [6]



Associate a jump sequence b_1^K with each sequence y_1^K
$P(y_1^K|x_1^K, u_1^K, K, e_1^I) = P(b_1^K|x_1^K, u_1^K, K, e_1^I) = \prod_{k=1}^K P(b_k|x_k, u_k)$



Inspired by work of Tillman

Input: x_1, x_2, $b \in \{0, +1, -1\}$

Output: y_1, y_2 with prob $(1-p)^2$
y_2, y_1 with prob p

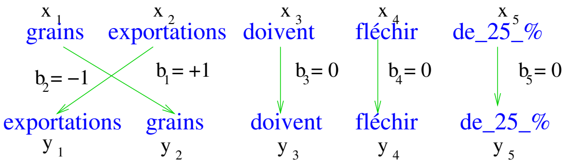
MJ-1 : maximum jump of 1

→ Properly parameterized

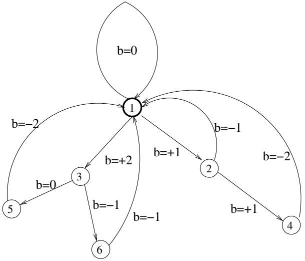
→ Not degenerate

[6] Kumar and W. Byrne (2005), Local phrase reordering models for statistical machine translation. Proc. HLT-EMNLP.

Phrase Swapping by WFSTs [6]



Associate a jump sequence b_1^K with each sequence y_1^K
$P(y_1^K|x_1^K, u_1^K, K, e_1^I) = P(b_1^K|x_1^K, u_1^K, K, e_1^I) = \prod_{k=1}^K P(b_k|x_k, u_k)$



Inspired by work of Tillman

Input: $x_1, x_2, b \in \{0, +1, -1\}$

Output: y_1, y_2 with prob $(1-p)^2$
y_2, y_1 with prob p

MJ-2 : maximum jump of 2

→ Properly parameterized

→ Not degenerate

[6] Kumar and W. Byrne (2005), Local phrase reordering models for statistical machine translation. Proc. HLT-EMNLP.

Incorporating Reordering in Translation with WFSTs

WFSTs can implement local phrase reordering

- Reodering prior to translation:
  - English phrases in French phrase order
  - Difficult to realize with WFSTs; can build permutation transducers [7]
- Tried the opposite approach, namely reordering after phrase translation:
  - Generate a French phrase sequence in English phrase order
  - Reorder sequence into French phrase order via Local Phrase Reordering Model
- Possible to realize with WFSTs both in alignment and translation
  - No English phrase reordering process
- Reordering is done prior to Insertion of Target Phrases
  - word alignments within phrases can span fairly long distances

Can perform embedded reestimation of reordering model parameters

- Phrase pair dependent reordering probability : $P(b_k|x_k, u_k)$
- Estimated via Viterbi approximation to EM
- Exact estimation: alignments are done under the translation model

[7] S. Kumar and W. Byrne. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. Proc. HLT-NAACL, 2003.

TTM and Phrase Reordering

Good gains from reordering in both Arabic→English and Chinese→English

- Jump-1 might be as good as Jump-2 (in this formulation)
- little gain in Chinese→English from parameter estimation

Translation under MJ-1 and MJ-2 reordering with a 4-gram LM.

	BLEU (%)					
Reordering	Arabic-English			Chinese-English		
Model	02	03	04	02	03	04
None	37.5	40.3	36.8	24.2	23.7	26.0
MJ-1 flat	40.4	43.9	39.4	25.7	24.5	27.4
MJ-1 VT	41.3	44.8	40.3	25.8	24.5	27.8
MJ-2 flat	41.0	44.4	39.7	26.2	24.8	27.8
MJ-2 VT	41.4	45.0	40.2	26.4	24.8	27.8

Influence of Language Model Order on Reordering in Translation

BLEU Over Merged Eval Sets (Eval02,Eval03,Eval04)

	BLEU (%)					
	A-E			C-E		
	2g	3g	4g	2g	3g	4g
None	21.0	36.8	37.8	16.1	24.8	25.0
MJ-1	23.4	40.4	41.6	16.2	25.9	26.5
MJ-2	23.3	40.4	41.6	16.0	26.0	26.8

These Simple Reordering Models Benefit from Better Language Models

Outline

Introduction

TTM – Phrase Translation via WFSTs

- WFST Operations
- TTM
- Phrase Pairs
- Phrase Segmentation
- Phrase Translation
- Phrase Insertion
- Phrase Reordering

TTM – Speech Translation

- A Generative Model of Speech Translation
- Translation of ASR Lattices

Alignment Models

- Model-4
- Word-to-Word HMM
- Word-to-Phrase Alignment, Phrase Pair Posteriors, Bigram Translation
- Translation Performance

Discussion

- Announcements
- Unsolicited Opinions

Statistical Phrase-Based Speech Translation [8]

Objective: Tight Integration of a Phrase-Based SMT System with ASR

- Unified statistical modeling framework

- Straightforward translation from speech without extensive reformation of the underlying statistical models or the ASR or SMT systems themselves

Problem: How to translate ASR lattices ?

[8] L. Mathias and W. Byrne (2006), *Statistical Phrase-Based Speech Translation*. ICASSP

A Generative Model for Text Translation

	Target Sentence		Target Phrases		Source Phrases		Source Sentence
	$\mathbf{T}$	$\longleftarrow$	$\mathbf{V}$	$\longleftarrow$	$\mathbf{U}$	$\longleftarrow$	$\mathbf{S}$
	Models $\longrightarrow$	$P(\mathbf{T} \mid \mathbf{V})$		$P(\mathbf{V} \mid \mathbf{U})$		$P(\mathbf{U} \mid \mathbf{S})$	$P(\mathbf{S})$
	FSMs $\longrightarrow$	Ω		Φ		W	G
		Target Phrase Mapping		Phrase Translation, Reordering		Source Phrase Mapping	Source Language Model

Ω : All foreign phrase sequences that could have generated the foreign text

Translation system translates phrase sequences, rather than word sequences

- This is done by first mapping the sentence into all its phrase sequences

A Generative Model for Speech Translation

	Target Speech		Target Sentence		Target Phrases		Source Phrases		Source Sentence
	A	$\longleftarrow$	$\mathbf{T}$	$\longleftarrow$	$\mathbf{V}$	$\longleftarrow$	$\mathbf{U}$	$\longleftarrow$	$\mathbf{S}$
Models →		$P(\mathbf{A} \mid \mathbf{F})$		$P(\mathbf{F} \mid \mathbf{V})$		$P(\mathbf{V} \mid \mathbf{U})$		$P(\mathbf{U} \mid \mathbf{E})$	$P(\mathbf{E})$
FSMs →		$\mathcal{L}$		Ω		Φ		W	G
		ASR Word Lattice		Target Phrase Mapping		Phrase Translation, Reordering		Source Phrase Mapping	Source Language Model

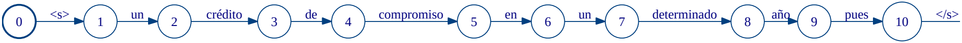
Target Phrase Segmentation is applied to the foreign language word lattice

$\mathcal{L} \cdot \Omega$: foreign phrase sequences that could have generated the foreign speech

Phrases are extracted from the ASR lattice, with scores, rather than from text

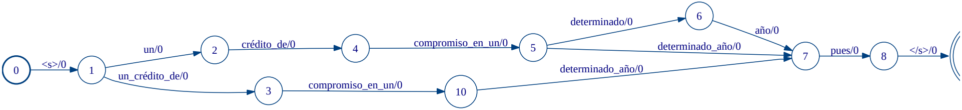
The translation system still translates phrase sequences and not word sequences

Target Phrase Segmentation – Translation of Text



Sentence Acceptor

⇓ Target Phrase Segmentation Transducer

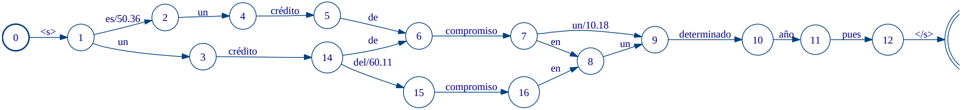


Phrase Sequence Lattice Ω

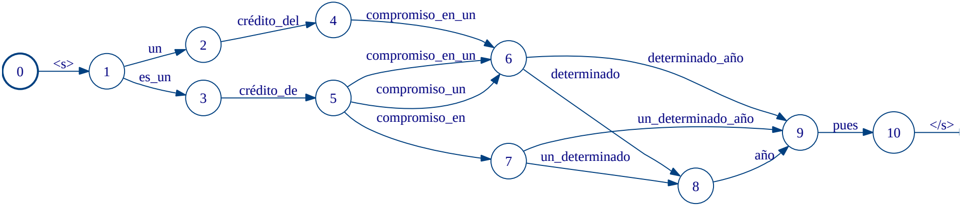
Phrase Sequence Lattice contains the phrase sequences in the foreign text

Target Phrase Segmentation – Translation of ASR Word Lattices

ASR Word Lattice



⇓ Target Phrase Segmentation Transducer



ASR Phrase Sequence Lattice

Phrase Sequence Lattice has the foreign phrase sequences in the speech lattice

- Phrase sequences correspond to the translatable word sequences in the lattice
- The lattice contains weights from the ASR system
- Translating this foreign phrase lattice is MAP translation of the foreign speech

Direct Translation of ASR Word Lattices

Original Problem: How to translate ASR lattices ?

New Problem: How to extract phrases from lattices ?

‘Speech Translation’ is recast as an ASR analysis problem in which the goal is to extract translatable foreign language phrases from ASR word lattices

Step 1. Perform foreign language ASR to generate a foreign language word lattice $\mathcal{L}$

Step 2. Analyze foreign language word lattices and extract phrases to be translated

Step 3. Build the target language phrase mapping transducer Ω

Step 4. Compose $\mathcal{L}$ and Ω to create the foreign language ASR Phrase Lattice

Step 5. Translate the foreign language phrase lattice using the TTM as if it were a phrase lattice extracted from a foreign language text sentence.

Tight coupling of ASR and SMT models and systems :

- Step 1 is a ‘standard’ ASR operation

- Steps 3, 4, and 5 are ‘standard’ TTM operations

Experimental Results

- Lattice translation should be no worse than translation from ASR transcriptions

- Prior to recently developed techniques[9] the ambiguity introduced by ASR lattices was reported to degrade translation performance

- Posterior-based pruning is used to control the amount of ambiguity presented to the translation system

Initial experiments in phrase-based speech translation using the TTM

TC-STAR 2005 Chinese-English Broadcast News Translation Task (BLEU)

	Mandarin Source	DEV	EVAL
Monotone Phrase Order	Ref. Transcription	16.1	18.8
	ASR 1-Best	14.8	13.6
	ASR lattice	15.0	13.8
MJ-1 VT Phrase Reordering	Ref. Transcription	16.1	19.3
	ASR 1-Best	15. 0	13.8
	ASR lattice	15.1	14.0

[9] For example, E. Matusov, S.Kanthak, H. Ney (2005), On the integration of speech recognition and statistical machine translation, InterSpeech.
N. Bertoldi and M. Federico (2005), A new decoder for spoken language translation based on confusion networks, ASRU.

Extracting Translatable Phrases from ASR Lattices

Modeling problems :

- Proper inclusion of the foreign language model (in progress ...)
- Phrase sequences extracted from ASR lattices might not appear in parallel text
  - ASR errors
  - spoken language phenomena, disfluencies, silences, ...
  - the usual genre mismatches
- The ASR and MT systems must be very compatible for this approach to work
- Possibly apply Rich Text transcription methods developed to ‘clean’ ASR output
  - Inserting SU markers, phrase boundaries, etc. into ASR lattices

Outline

Introduction

TTM – Phrase Translation via WFSTs

- WFST Operations
- TTM
- Phrase Pairs
- Phrase Segmentation
- Phrase Translation
- Phrase Insertion
- Phrase Reordering

TTM – Speech Translation

- A Generative Model of Speech Translation
- Translation of ASR Lattices

Alignment Models

- Model-4
- Word-to-Word HMM
- Word-to-Phrase Alignment, Phrase Pair Posteriors, Bigram Translation
- Translation Performance

Discussion

- Announcements
- Unsolicited Opinions

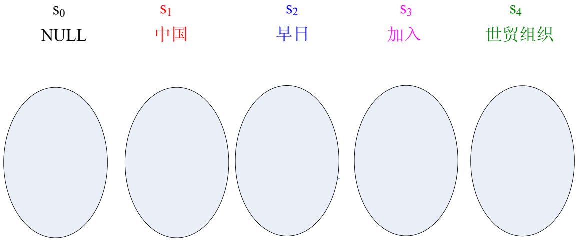
IBM Model-4 for Word Alignment (Review)

Model 4 is a powerful model of word alignment within sentence pairs

Features:

- Lexical translation model: as in Model 1 and HMM alignment model
- Fertility: probability distribution over the number of target words each source word can generate
  - an approximation to *phrase modeling*
- NULL Translation Model: allows target words to be generated without a corresponding source on the source side
- Distortion Model: describes how target words are distributed throughout the target sentence when generated from a single source word

Model-4 Generation: Step 1 – Tablets



s_0 NULL

s_1 中国

s_2 早日

s_3 加入

s_4 世贸组织

Create a tablet for each source word

Model-4 Generation: Step 2 – Fertility

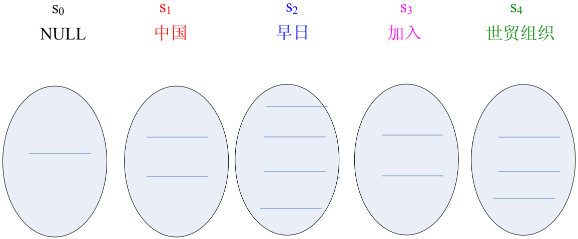
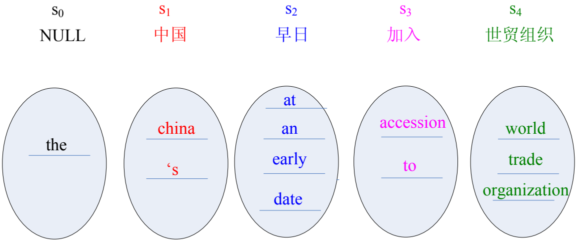


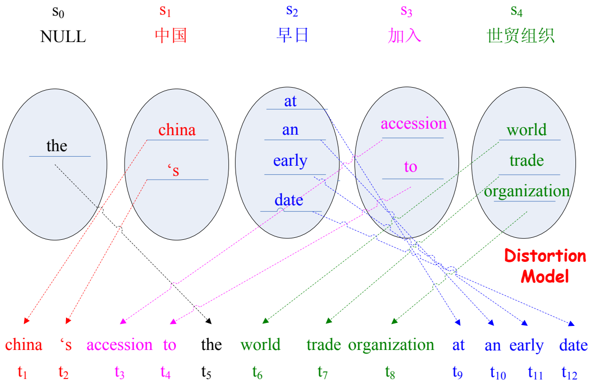
Table lookup to decide fertility: # of target words connected

Model-4 Generation: Step 3 – Fill in Tablet Positions



Sample target words from translation table i.i.d.

Model-4 Generation: Step 4 – Fill in the English Sentence



IBM Model-4 Summary

Very powerful model of word alignment in translated sentences

- Fertility and Distortion begin to capture generation of phrases in translation.

Model-4 has several shortcomings

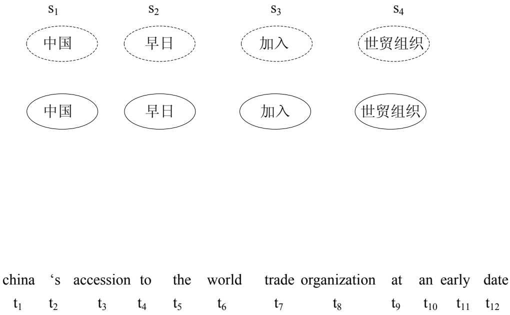
- Deficiency - The Distortion Model used to move words from the tables into position in the sentence may place two words in one sentence position. This is a drawback – some probability mass under the model is assigned to what are effectively non-sentences. The probabilities over alignments and generated sentences do not sum to 1.0 . The models are therefore called deficient.
- Parameter estimation and word alignment are difficult to implement
    - Unlike Model-1, Model-2, and HMM Alignment, dynamic programming based algorithms are not available for Model-4.

  Difficult to balance modeling power & computational efficiency in alignment.

Despite these difficulties, IBM Model-4 can be used to generate high-quality word alignments over parallel text, especially as implemented under the GIZA++ toolkit[10].

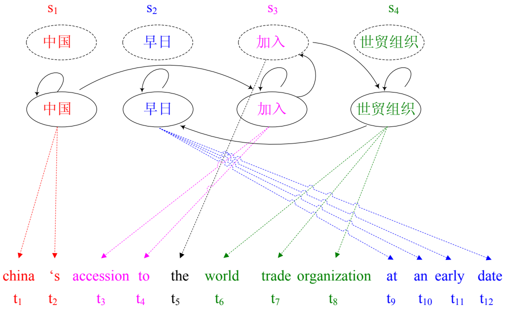
[10] http://www.fjoch.com/GIZA++.html

Word-to-Word HMM Alignment [11]



[11] S. Vogel, H. Ney, C. Tillmann (1996), HMM Based Word Alignment in Statistical Translation. COLING.
F.J. Och & H. Ney (2003), A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics.

Word-to-Word HMM Alignment

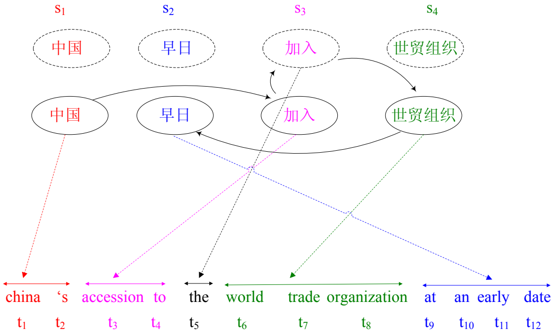


- A state sequence specifies a word-to-word alignment
- Words are generated one by one
- One transition emits one target word

Can HMM alignment be made as powerful as IBM-4 alignment ?

- Investigate modeling methods originally developed for ASR

Word-to-Phrase HMM Alignment [13]



- Target phrases rather than words are emitted at each state transition
- A state sequence specifies a word-to-phrase alignment
- An instance of a segmental HMM, similar to those developed for ASR [12]
  - Efficient alignment and parallel estimation procedures are possible

[12] M. Ostendorf, V. Digalakis, and O. Kimball (1996), From HMMs to segment models: a unified view of stochastic modeling for speech recognition. IEEE Trans. Acoustics, Speech and Signal Processing.

[13] Y. Deng and W. Byrne (2005), HMM word and phrase alignment for statistical machine translation. Proc. HLT-EMNLP.

Bigram Translation Probabilities

Goal: replace weak i.i.d. word-to-word translation [13]

- Add context dependence into translation, while respecting HMM dependencies

$$P(\text{world trade organization}|f = \text{世贸组织}; 3) = ?$$

$$= t(\text{world}|\text{f}) \cdot t(\text{trade}|\text{f}) \cdot t(\text{organization}|\text{f}) \Longleftarrow \text{i.i.d.}$$

$$= t(\text{world}|\text{f}) \cdot t_2(\text{trade}|\text{world}, \text{f}) \cdot t_2(\text{organization}|\text{trade}, \text{f}) \Longleftarrow \text{bigram}$$

Model	i.i.d.	bigram
$P(\text{world}|\text{世贸组织})$	0.06	0.06
$P(\text{trade}|\text{world}, \text{世贸组织})$	0.06	0.99
$P(\text{organization}|\text{trade}, \text{世贸组织})$	0.06	0.99
$P(\text{world trade organization}|\text{世贸组织}, 3)$	0.0002	0.0588

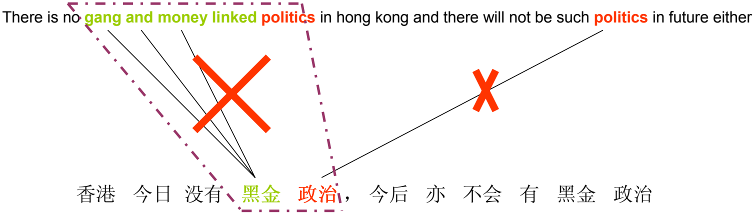
- Estimated in the same way as bigrams in monolingual LMs
- Data sparseness requires smoothing – Witten-Bell backoff

[13] Y. Deng and W. Byrne (2005), HMM word and phrase alignment for statistical machine translation. Proc. HLT-EMNLP.

A. L. Berger, S. Della Pietra, and V. J. Della Pietra (1996) A maximum entropy approach to natural language processing. Computational Linguistics

Avoiding Viterbi Alignments in Extracting Phrase Pairs

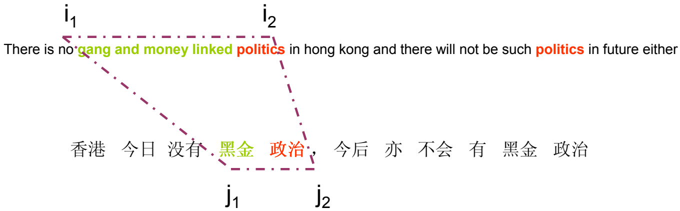
Word alignments are not perfect


There is no gang and money linked politics in hong kong and there will not be such politics in future either
香港 今日 没有 黑金 政治 ， 今后 亦 不会 有 黑金 政治

- Relying on the one-best word alignment may exclude some valid phrase pairs
- Goal : define a probability distribution over phrase pairs
- Allow more control over the generation of phrase pairs

Model-Based Phrase pair Posterior Distributions

Avoid relying on a single one-best alignment


i_1
i_2
There is no **gang and money linked** **politics** in hong kong and there will not be such **politics** in future either
香港 今日 没有 **黑金** **政治**， 今后 亦 不会 有 黑金 政治
j_1
j_2

- Define a set of alignments that align words to words within phrases

$$A(i_1, i_2; j_1, j_2) = \{ \, a_1^m \ : a_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2] \, \}$$

- Calculate the posterior probability that the source phrase and the target phrase are aligned given the two sentences

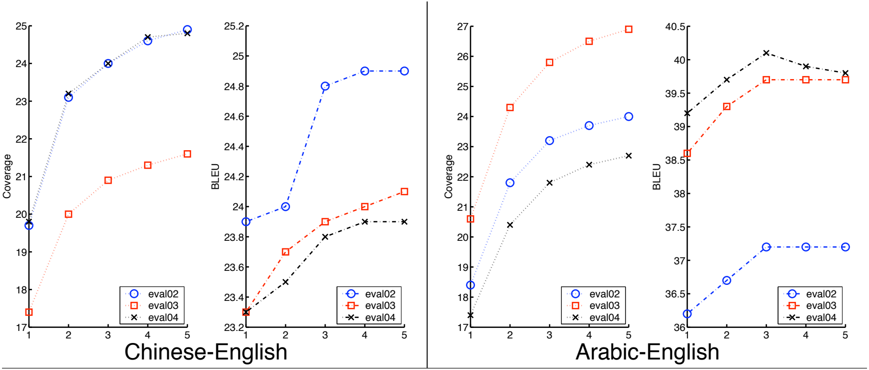
$$P(\, A(i_1, i_2; j_1, j_2) \mid t, s) = P(t, A(i_1, i_2; j_1, j_2)|s) \ / \ P(t|s)$$

- Easily found under the Word-to-Phrase HMM
- Cannot be found exactly under IBM Model-4

Phrase pair Augmentation Under the Posterior Distribution [14]

Translation Systems : Chinese-English FBIS and Arabic-English News

- Start with the phrase pairs extracted from the Viterbi alignments
- Add phrase pairs based on the posterior distribution
  - Keep all phrase pairs whose posterior is above a threshold



Chinese-English

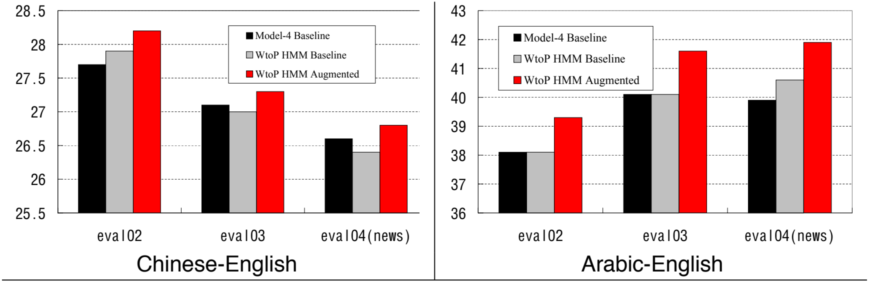
Arabic-English

- Balance coverage against phrase translation quality

[14] Y. Deng and W. Byrne (2005), HMM word and phrase alignment for statistical machine translation. Proc. HLT-EMNLP.

TTM Performance with Word-to-Phrase HMM Alignments :
- Comparison to IBM-4 and Phrase Pair Augmentation

Baseline JHU/CU system development for NIST 2005 MT Eval



- Used all parallel text available from LDC as of June 2005
  - C-E : 200M English words (FBIS, Xinhua, HK News, ..., UN)
  - A-E : 130M English words (News sources, UN)

Outline

Introduction

TTM – Phrase Translation via WFSTs

- WFST Operations
- TTM
- Phrase Pairs
- Phrase Segmentation
- Phrase Translation
- Phrase Insertion
- Phrase Reordering

TTM – Speech Translation

- A Generative Model of Speech Translation
- Translation of ASR Lattices

Alignment Models

- Model-4
- Word-to-Word HMM
- Word-to-Phrase Alignment, Phrase Pair Posteriors, Bigram Translation
- Translation Performance

Discussion

- Announcements
- Unsolicited Opinions

Towards Integrated Alignment and Decoding

Alignment: Given a source sentence S and a target sentence T, construct a joint distribution over their alignment.

$$P(S,\mathcal{A},T) = \underbrace{P(T|\mathcal{A},S)}_{\text{Translation Model}}\ \underbrace{P(\mathcal{A}|S)}_{\text{Alignment Model}}\ \underbrace{P(S)}_{\text{Language Model}}$$

Translation (ideal) : Given T, find a translation $\widehat{S}$ and an alignment $\widehat{\mathcal{A}}$ as

$$(\,\widehat{S},\widehat{\mathcal{A}}\,) = \operatorname*{argmax}_{S,\mathcal{A}}\ P(T|\mathcal{A},S)\,P(\mathcal{A}|S)\,P(S)$$

Translation (current reality) : lacks integrated modeling and decoding

- Models are trained & alignments are generated over the training set
  - models are not applied outside the training set – no expectation of generalization
- The models are discarded, and the alignments are kept
- PPI, etc. are extracted from the alignments and used in translation

Goal: Embedded Model Estimation

- same models for alignment and translation – 'what works' for ASR
- needed for : MMI, clustering, context dependence, ...

Simple and Efficient Models for Alignment and Translation

Simple and Efficient: efficient, dynamic programming procedures for parameter estimation and decoding (translation)

- Sacrifice elegance for computational efficiency
  - Word-to-Phrase HMMs, with extensions, yield alignments of comparable quality to those produced by IBM Model-4
- Make maximum use of all available parallel text
  - Phrase pair induction to increase phrase coverage of the test set
  - Fast parallelizable training with EM-based HMM estimation procedures
- Extensions respect the underlying conditional independence assumptions
  - Extending the TTM from text translation to speech translation is (fairly) straightforward given the underlying generative model
  - HMM alignment extensions may add complexity but do still allow DP-based estimation and alignment procedures
- Have developed a translation framework within which it should be possible to apply the HMM refinements associated with state of the art ASR

IEEE Transactions on Audio, Speech and Language Processing

Papers Welcome:

IEEE Transactions has expanded its scope to include language processing

- Submissions on Translation, and Speech Translation in particular, are invited

Positions Available at the CUED Machine Intelligence Lab

AGILE: Autonomous Global Integrated Language Exploitation

Research Associates in Speech and Language Processing

Research Studentships in Speech and Language Processing

Currently accepting applications. Successful candidates will be expected to contribute primarily to ongoing research in the following areas:

- development of statistical machine translation systems;
- integration of statistical machine translation and large vocabulary speech recognition systems.

See http://mi.eng.cam.ac.uk/jobs/

New! – TTM Tutorial is Available

German→English translation based on

- Europarl corpus
- Giza++ alignments
- AT&T FSM Toolkit
    - any FSM toolkit should work ...

Tutorial steps through building and using all the transducers

Very much an alpha version, but available to anybody who's interested

New! – MTTK : Machine Translation Toolkit

MTTK is a collection of software tools for the alignment of parallel text for use in Statistical Machine Translation. The toolkit was written by Yonggang Deng in the course of his Ph.D. at The Johns Hopkins University Center for Language and Speech Processing.

With MTTK you can ...

- Align document translation pairs at the sentence or sub-sentence level.
- Train statistical models for parallel text alignment. The following are supported :
  - IBM Model-1 and Model-2
  - Word-to-Word HMMs
  - Word-to-Phrase HMMs, with bigram translation probabilities
- Parallelize your model training procedures.
- Generate word-to-word and word-to-phrase alignments of parallel text.
- Extract word-to-word translation tables from the models and the aligned bitext.
- Extract phrase-to-phrase translation tables from aligned parallel text.
- Use the HMM alignment models to induce phrase translations.
- Edit the C++ sources to implement new estimation and alignment procedures.

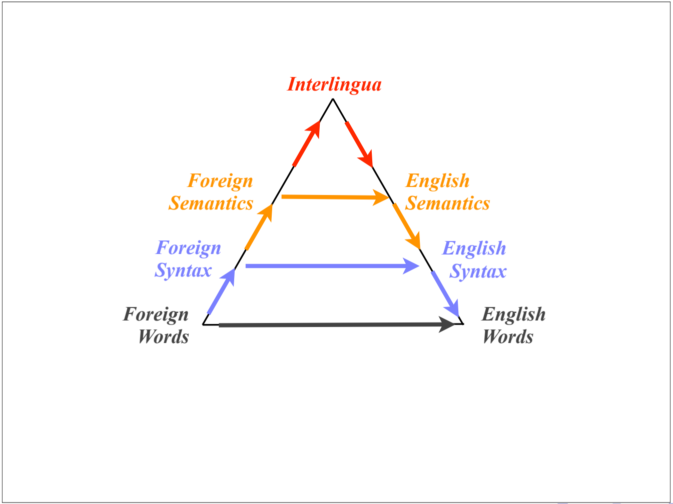
See http://mi.eng.cam.ac.uk/∼wjb31/distrib/mttkv1/

Version 1.0 - Yonggang Deng, Bill Byrne

The Machine Translation Pyramid

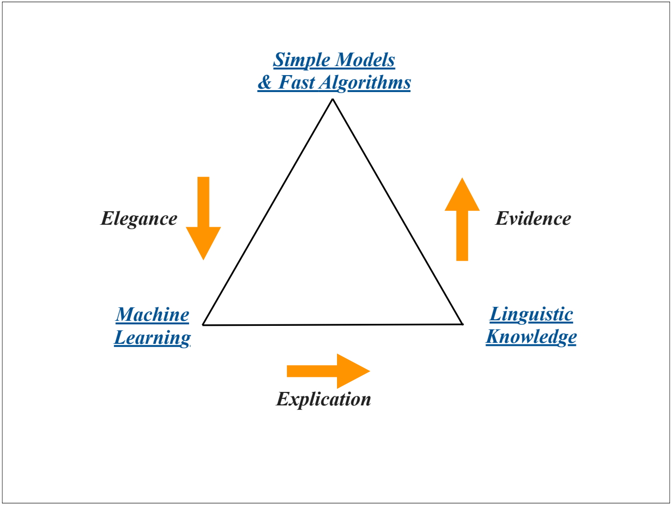
Casts the problem in familiar terms

- strengths and weaknesses of the formulation are obvious



The *Statistical* Machine Translation Pyramid

Building good systems requires balancing competing concerns



Simple Models & Fast Algorithms

Elegance

Evidence

Machine Learning

Linguistic Knowledge

Explication

Thanks!