Engineering Part IIB: Module 4F11

Speech and Language Processing

Lecture 13: Statistical Machine Translation Systems

Bill Byrne

Lent 2016

Cambridge University Engineering Department

Automatic Measurement of Translation Quality

Automatic performance metrics have been central to the development of large statistical language processing systems

- Word/Character Error Rate (WER): ASR , OCR, ...
- other metrics for other tasks : precision/recall, ...

These are all relative to *human performance* over defined test sets

- human translations are obtained *once* over a fixed test set
- system performance can be measured *many times* relative to :
  - human performance, directly
  - performance of other systems, indirectly
- automatic metrics make incremental system improvement possible
- metrics can be incorporated into estimation and decoding algorithms

Automated evaluation metrics for Machine Translation (as for other tasks) should

- be inexpensive to compute
- require no human participation
- be closely correlated with human perception of translation quality

Automatic Measurement of Machine Translation Quality

BLEU[1] is an MT metric based on n-gram precision

- An example of computing Bleu against a single reference translation:

Reference : mr. speaker , in absolutely no way .
Hypothesis : in absolutely no way , mr. chairman .

BLEU Computation

n-gram matches				BLEU
1-word	2-word	3-word	4-word	$\left(\frac{7}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5}\right)^{\frac{1}{4}} = 0.3976$
7/8	3/7	2/6	1/5	

- Can be generalized to multiple references
- Typically also includes a length penalty
- Correlates well with human judgments of translation
- By comparing automatic translations to human references we obtain implicit measurements of fluency and accuracy

[1] Papineni, K., Roukos, S., Ward, T., & Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Pages 311318 of: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).

An Example of Good Translation

Original Chinese Gloss

(By) (2005 year) (internet) (,) (whole country) (users) (will) (reach) (0.2 billion)

Automatic Translation

By 2005 , the number of internet users will reach 200 million

Human References

By 2005 , the number of internet users in the whole country will reach 200 million
By 2005 , the number of internet users in China is estimated to be 200 million
By 2005 , internet customers across the country is to reach 200 million
In 2005 , the internet users in China will total 0.2 billion

- The translation agrees fairly closely with the reference translations
- The translation is close to fluent
- Note the variability in the four reference translations.

An Example of Relatively Poor Translation

Original Word Segmented Chinese Pinyin (with tones) and gloss

sui1ran2 bei3feng1 hu1xiao4 , dan4 tian1kong1 yi1ran2 shi2fen1 qing1che4
(Although) (northern wind) (howl) (,) (but) (sky) (very) (clear)

Automatic Translation

Although wind howl . but the skies remain very tender

Human References

Although a north wind was howling , the sky remained clear and blue .
However , the sky remained clear under the strong north wind .
Despite the strong northerly winds , the sky remains very clear .
The sky was still crystal clear , though the north wind was howling .

- some resemblance to the reference translations
- not fluent
- questionable accuracy

BLEU Assigns High Scores to Good Translations

Example translations at various levels of translation performance as measured by the sentence-level BLEU score.

Translations	BLEU (%)
	60 – 70%
Afghan Earthquake Victims begin to rebuild their homes .	66.1
Prior to this , the ANC has issued a statement calling for the international community to respect the choice of the people and help them survive .	66.0
Statistics show that since 1992 , a total of 204 UN personnel have been killed , but only 15 criminals have been arrested .	64.4
Chavez emphasized that Venezuela needs peace , stability and reason for all parties should make joint efforts to end the conflict .	62.2
London Financial Times Index Friday at closing newspaper 5,292.70 points , up 31.30 points .	61.0

readable and fairly plausible

BLEU Assigns Middling Scores to Middling Translations

	$20-30\%$
Japan to temporarily freeze asked Russia to provide humanitarian assistance ,	30.0
Opposition Senator held that the president should focus more on domestic affairs and not eager to go abroad .	26.2
Taiwan DPP Legislator Chen Kim de fisheries groups to visit to Beijing .	23.9
Recently , the international community for the recent conflict , the fiercest Jenin camp conflict investigation of spreading .	20.8
opinion maintained : Gusmao victory is a strong possibility because he is considered the East Timor independence hero .	20.0

- probably misleading with respect to details
- contains readable sections

BLEU Assigns Low Scores to Poor Translations

	$0-10\%$
Japan Telecom company in 2000 to spend 5.5 billion dollars buy back .	0.0
However , the voting result shows that Zhu because there is no reason to be losing power by NPC deputies desolate .	0.0
77 private manufacturing enterprises also reported a foreign trade management right .	0.0
Identification Department found that college students of the certificate , many of them were fake .	0.0
The European Union would be implemented in steel imports temporary protective measures to discuss with the Chinese side ,	0.0
Georgia from a section of the great mountains Canyon withdrawal ,	0.0

barely readable and probably misleading

Calculation of BLEU

The goal is to calculate the BLEU score of a set of automatic translations $\{E^i\}_{i=1}^R$ against a set of reference translations, e.g. $\{E^i_{(1)}, E^i_{(2)}, E^i_{(3)}, E^i_{(4)}\}_{i=1}^R$.

- Set N to be the order of the highest n-gram to be considered – default is N=4
- For each sentence i, and for $n = 1, \ldots, N$, gather the following n-gram counts:
  - c_n^i : the number of hypothesized n-grams
  - $\bar{c}_n^i$: the number of correct n-grams, where the contribution of each distinct n-gram is *clipped* to the maximum number of occurrences in any one reference

  Example:

Hypothesis:	the the the the the the the
Reference 1:	the cat is on the mat
Reference 2:	there is a cat on the mat

  In this example, $c_1 = 7$, but $\bar{c}_1 = 2$
- Compute the precision for each n-gram order, $n = 1, \ldots, N$: $p_n = (\sum_i \bar{c}_n^i) / (\sum_i c_n^i)$
- Calculate the Brevity Penalty
  - Compute the shortest reference length : $r = \sum_i \min\{|E^i_{(1)}|, |E^i_{(2)}|, |E^i_{(3)}|, |E^i_{(4)}|\}$
  - Compute the hypothesis length : $c = \sum_i |E^i|$

$$BP = \begin{cases} 1 & c > r \\ \exp\left(1 - \frac{r}{c}\right) & c <= r \end{cases}$$

- The BLEU score is

$$\text{BLEU} = BP * \exp\{\sum_{n=1}^{N} \frac{1}{N} \log p_n\}$$

BLEU Can Be Used For SMT Development

BLEU is not an absolute measure of translation performance

- unlike Word Error Rate used in speech recognition
- most useful in indicating the relative quality between two different systems

Assessment using BLEU can be trusted when :

- many different translation systems are developed under BLEU, and results throughout development are published on standard test sets
- performance is frequently validated against human judgments

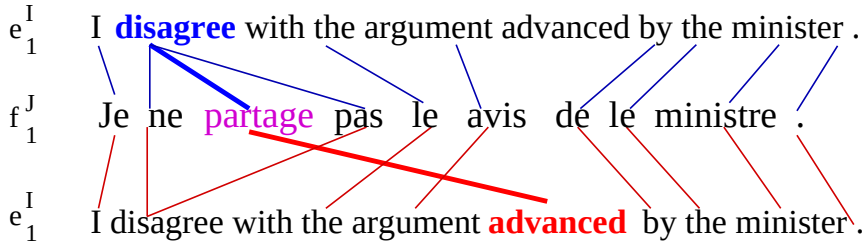
Current test sets used in the NIST MT evaluations

- $\sim 1K$ sentences / $\sim 25K$ words,
- four independently produced reference translations

There is extensive effort in improving/extending/replacing BLEU

Measuring Automatic Word Alignments Against Manual References

This figure shows two sets of alignments between a French sentence and an English sentence (the English sentence is shown twice).



- (top) B : automatic word alignments ← *produced by an alignment model*
- (bottom) B': reference word alignments ← *created by humans*

Alignment Error measures the number of non-NULL word alignments by which the automatic word alignment differs from the reference word alignment.

Step 1. Remove the NULL word links from B' and B to form $\bar{B}'$ and $\bar{B}$

Step 2. Compute $AE(B, B')$:

$$AE(B, B') = \frac{|\bar{B}'| + |\bar{B}| - 2\,|\bar{B} \cap \bar{B}'|}{|\bar{B}'| + |\bar{B}|} = \frac{10 + 10 - 2 * 9}{10 + 10} = \frac{2}{20} = 10\%$$

Alignment Error has been found to be a good, if not perfect, indicator of the quality of word alignment models: large reductions in alignment error are often correlated with improvements in translation quality. Alignment Error is often used as an intermediate quality measure in translation system development.

There is extensive effort in improving/extending/replacing Alignment Error

Phrase-to-Phrase Translation Models

Models of word-to-word translation are very useful for alignment of pairs of known sentence translations. However, word-to-word translation models do not capture the context that conveys syntactic and semantic information.

To avoid the limitations of word-based models, phrase-based translation segments a sentence to be translated into sequences of phrases which are then translated as entities.

Word alignment models are typically not used in translation

- They allow too much freedom in translation and reordering
- Translation (as opposed to alignment) requires more constrained models
- In this context, phrases are simply word sequences extracted from sentences.
- A phrase is a sequence of words which can be translated

Phrase-Based translation proceeds as follows :

- The sentence to be translated is segmented into foreign phrases. There are usually many possible phrase segmentations of the sentence.
- English phrases are extracted from word-aligned parallel text for use as potential translations
- The foreign phrases sequences are translated into English phrase sequences based on the phrases found in the parallel text.

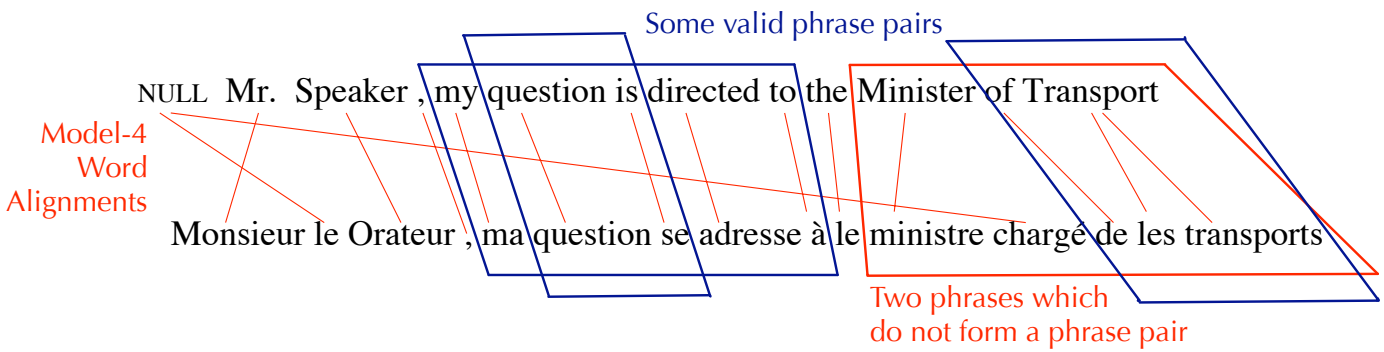
In phrase-based translation, word alignment models are used primarily for word-level alignment of the parallel text which is then used in extracting phrase translations

Finding Phrase Translations in Word-Aligned Parallel Text [2]

- Phrase Translations, called phrase pairs, are extracted from word-level alignments, typically generated using IBM-4 on the same parallel text over which it was trained
- Phrase Pairs cover patterns of word alignments in the training bitext
- Rules are defined to specify valid phrase pairs, for example:

Two phrases are aligned if their words align only with each other

- As an example, there are many possible phrase pairs which can be extracted from this word-aligned sentence pair under this rule:



However, some reasonable phrase pairs are excluded due to flawed word alignments

- Phrase translation probabilities are found by counting phrase pairs in the parallel text

[2] F. Och. 2002. *Statistical Machine Translation: From Single Word Models to Alignment Templates.* Ph.D. thesis, RWTH Aachen, Germany.

Phrase-Based SMT - A Simple Example (without probabilities)

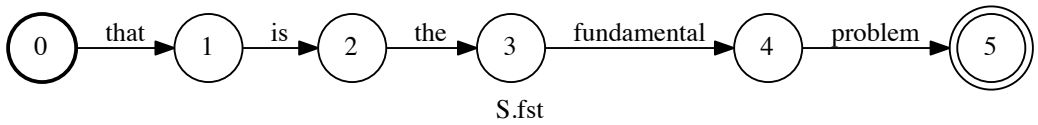
Source sentence: That is the fundamental problem

Phrase table: automatically extracted from automatically aligned parallel text

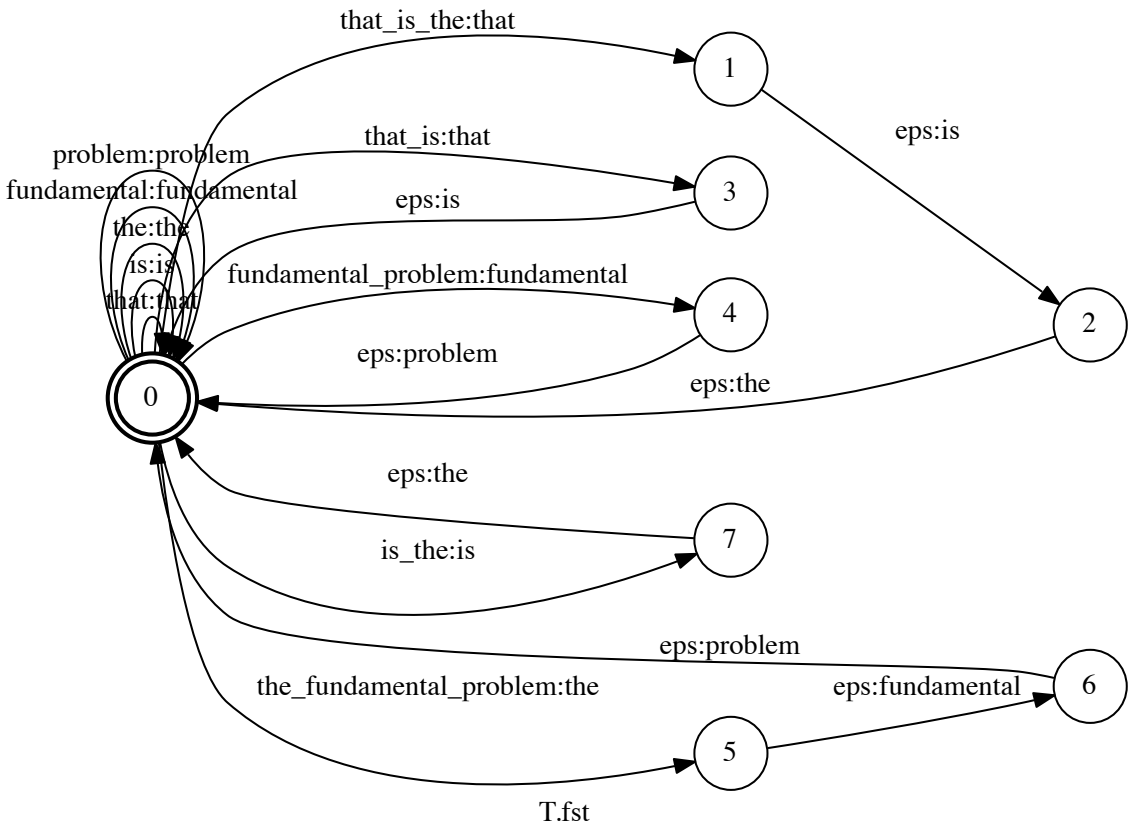
English	French
that_is_the	voila
these_are_the	voila
these_are	voila
those_are	voila
that_is_the	qui_est
that_is_the	qui_est_le
fundamental_problem	probleme_fondamental
the_fundamental_problem	probleme_fondamental
fundamental_problem	le_probleme_fondamental
fundamental_problem	la_difficulte_fondamental
the_fundamental_problem	le_probleme_fondamental
who_is	qui_est
it_is	qui_est_le
it_is_the	qui_est_le

Goal: a translation of the source sentence using the phrase table

Phrase-Based SMT - 1. Build a **source sentence acceptor** and a **source phrase segmentation transducer**

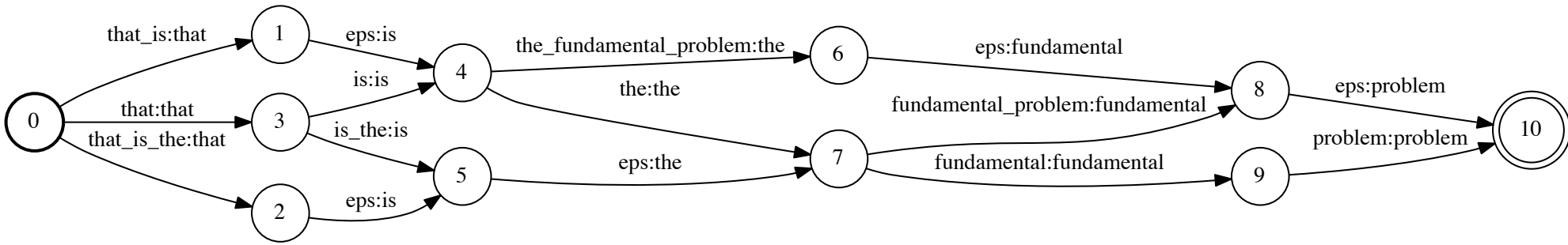


S.fst



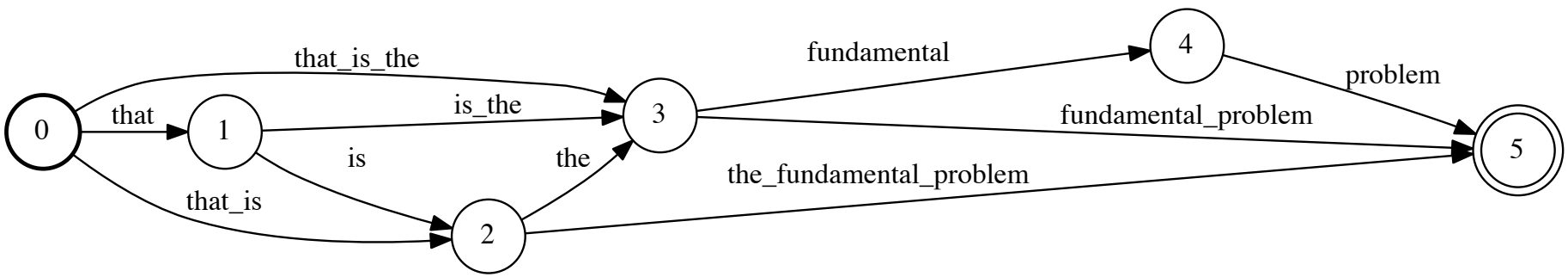
T.fst

Phrase-Based SMT - 2. Generate the **source phrase acceptor** via Composition and Projection



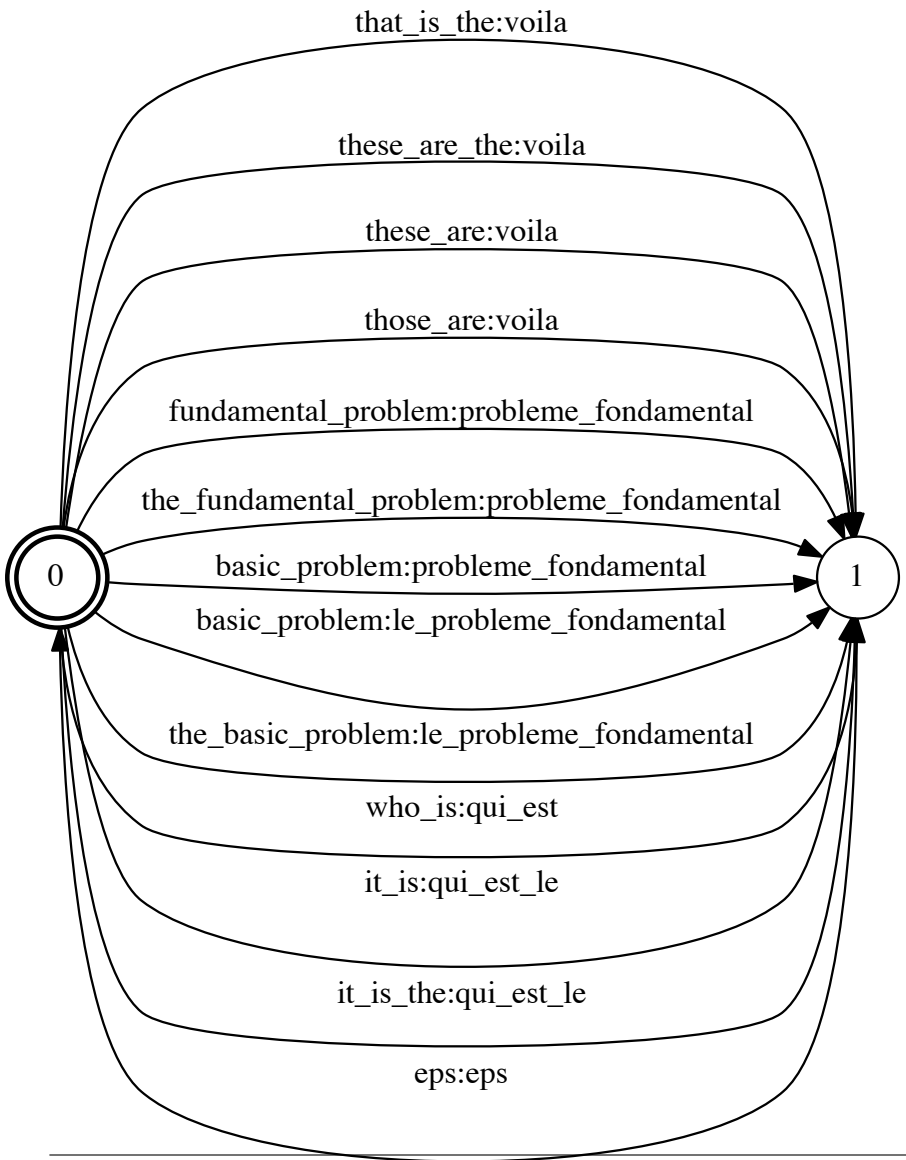
Source Word To Source Phrase.fst

```
fstcompose T.fst S.fst | fstproject | fstrmepsilon | fstminimize - U.fst
```



U.fst

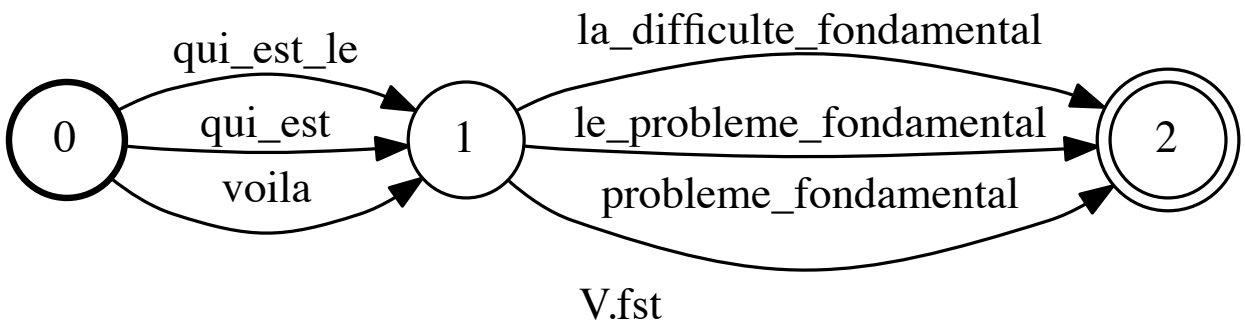
Phrase-Based SMT - 3. Build a **phrase translation transducer**



English	French
that_is_the	voila
these_are_the	voila
these_are	voila
those_are	voila
that_is_the	qui_est
that_is_the	qui_est_le
fundamental_problem	probleme_fondamental
the_fundamental_problem	probleme_fondamental
fundamental_problem	le_probleme_fondamental
fundamental_problem	la_difficulte_fondamental
the_fundamental_problem	le_probleme_fondamental
who_is	qui_est
it_is	qui_est_le
it_is_the	qui_est_le

Phrase-Based SMT - 4. Apply the **phrase translation transducer** to the **source phrase acceptor** via Composition to generate the **target language phrase acceptor**

```
fstcompose U.fst Y.fst | fstproject --project_output - V.fst
```



- This yields French phrases in English phrase order.
- Need to include context and reordering (a.k.a movement)

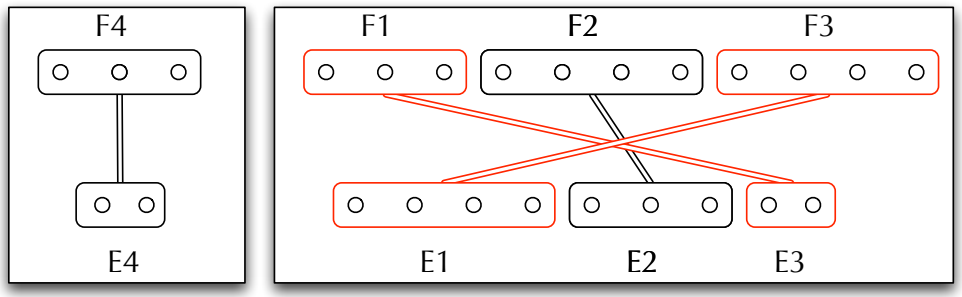
Hierarchical Phrase-Based Translation

Goal: synthesise complex translation rules from phrase patterns observed in the parallel text

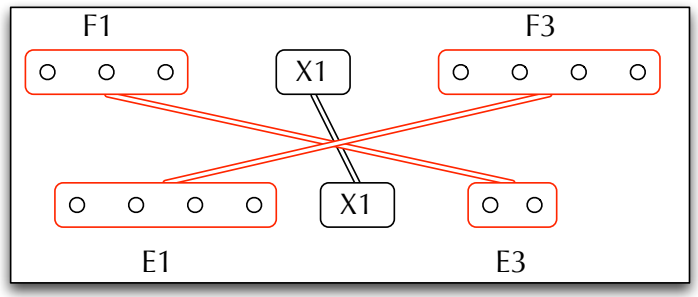
- Examples of phrasal translation rules we might extract from aligned parallel data

`ne X1 pas` → `not X1`
`le X2 X1` → `X1 of the X2`

Phrase-to-Phrase Translation Rules Found in the Parallel Text



Hierarchical Phrase-based Translation Rule



Can now translate `E1 E4 E3` as `F1 F4 F3`

- Extract translation rules with **non-terminals** (X's)
- Non-terminals allow other rules to be applied
  - Phrase translation rules:
    `E4` → `F4`
    `E1` → `F3`
    `E2` → `F2`
    `E3` → `F1`
  - Hierarchical phrase translation rule:
    `E1` **X1** `E3` → `F1` **X1** `F3`
- rules also have **weights**, e.g. negative log probabilities (omitted here for clarity)

RTN Translation Representations

Individual automata represent applications of translation rules across **source spans**
The example below shows some translations under a single source phrase segmentation:

- T[1,4] represents rules that can span source source sentence positions 1 through 4
- The entire set of translations is expressed by T[1,8]

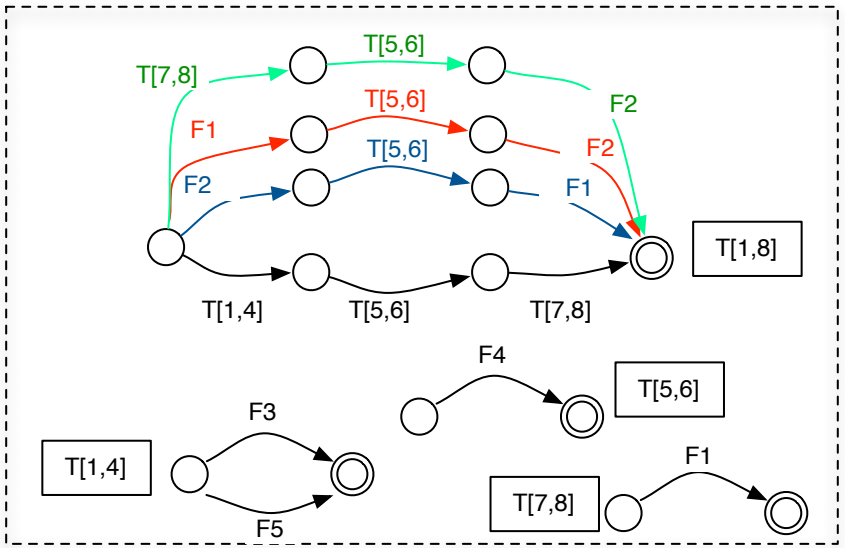
Phrase translation rules:

E1 → F3
E1 → F5
E4 → F4
E3 → F1

Hierarchical phrase translation rules:

E1 X1 E3 → F1 X1 F2
E1 X1 E3 → F2 X1 F1
E1 X1 X2 → X2 X1 F2

- E1 covers 1-4
- X1 covers 5-6
- X2 covers 7,8



- The arc weights contain the translation rule scores (negative log probabilities)
- After the RTN is built, it is **expanded** into an equivalent WFSA
- Language models are applied via composition
- The best translation is found by the `fstshortestpath` operation

Language Models for SMT - Back-off N-Gram Models (review)

Recall the general form for a back-off n-gram language model:

- The LM history is the most recent n-1 words

$$P(e_1^K) = \prod_{k=1}^{K} P(e_k|e_1^{k-1}) \approx \prod_{k=1}^{K} P(e_k|e_{k-n+1}^{k-1})$$

- The maximum likelihood estimate of $P(e_k|e_{k-n+1}^{k-1})$ is based on counts $f(\cdot)$ collected over collections of monolingual text

$$P(e_k|e_{k-n+1}^{k-1}) = \frac{f(e_{k-n+1}^{k})}{f(e_{k-n+1}^{k-1})}$$

- As the length of the history grows, the counts of many n-grams will be zero. Rather than assign zero probability to word sequences, backing off and discounting are applied

$$P(e_k|e_{k-n+1}^{k-1}) = \begin{cases} d(f(e_{k-1n+1}^{k}))\frac{f(e_{k-n+1}^{k})}{f(e_{k-n+1}^{k-1})} & f(e_{k-n+1}^{k}) > C \\ \alpha(e_{k-n+1}^{k-1})P(e_k|e_{k-n+2}^{k-1}) & \text{otherwise} \end{cases}$$

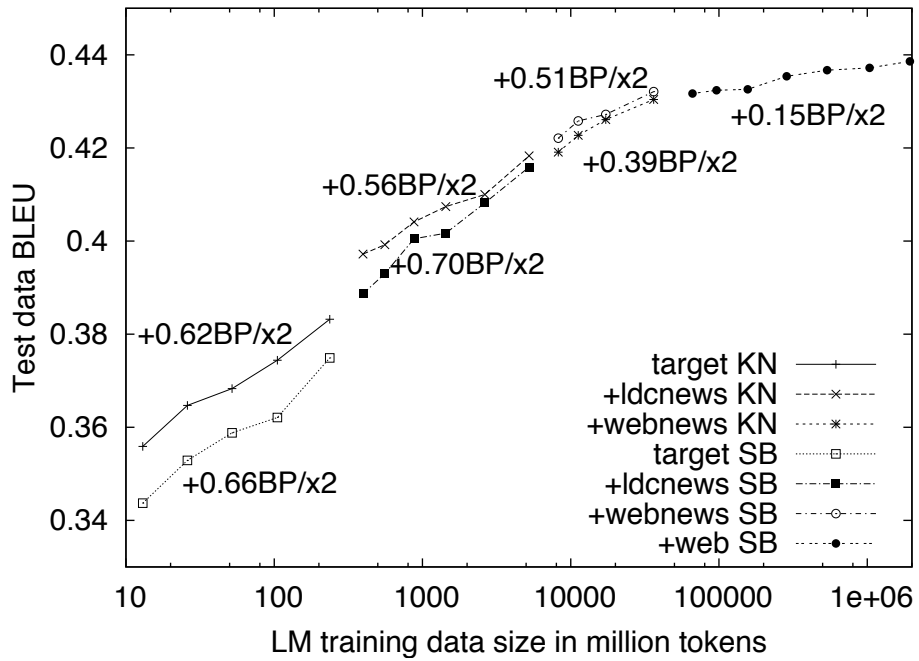
- discount weights (d) and back-off weights (α) are estimated, e.g. by Kneser-Ney, so that a proper probability distribution is defined over all sequences

Language Models for SMT - Zero Cutoff, Stupid Back-Off Models

Problem: Backoff N-Gram parameter estimation is difficult as training sets grow
Solution: Don't estimate, just count [3]

Stupid Backoff

$$S(e_i|e_{i-n+1}^{i-1}) = \begin{cases} \frac{f(e_{i-k+1}^{i})}{f(e_{i-k+1}^{i-1})} & \text{if } f(e_{i-k+1}^{i}) > 0 \\ \alpha\, S(e_i|e_{i-k+2}^{i-1}) & \text{otherwise} \end{cases}$$



Avoids the need to estimate discounting and back-off parameters, however two problems arise:

- Stupid back-off procedure assigns *scores* to sequences, not probabilities
- WFSAs with epsilon arcs used in an approximate implementation does not work well here

[3] T. Brants et al. 2007. Large Language Models in Machine Translation. EMNLP