Hierarchical Phrase-based Translation with Weighted Finite-State Transducers

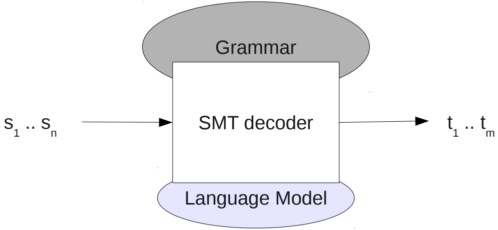
Gonzalo Iglesias

Machine Intelligence Laboratory

Cambridge University Engineering Department

Leuven, 31st of May 2011

Exactness in Search I



- The search space is the intersection of s_1^n with the grammar.
- In order to improve performance we may consider exploring bigger search spaces
- But if it is too big for the decoder, local pruning is needed.

Exactness in Search II

- Local pruning leads to search errors
- CUED Machine Translation group: Exactness and WFSTs.
  - Phrase-based Translation: TTM[1]
  - HiFST + shallow-N grammars

[1] Kumar, Shankar, Yonggang Deng, and William Byrne. 2006. **A weighted finite state transducer translation template model for statistical machine translation.** Natural Language Engineering.

This talk

- Hierarchical Phrase-based Translation
- HiFST: Hiero decoder using WFSTs
- Shallow-N grammars
- Results on several Translation Tasks
  - Contrast with Hiero Cube pruning decoder for AREN and ZHEN
  - HiFST and LMBR (rescoring and system combination)
  - WMT10 (FR/EN and SP/EN tasks)

Hierarchical Phrase-based Translation

- Hierarchical grammars and decoders first introduced by Chiang[2]
- Hierarchical rules (phrases with gaps) allow generalization and reordering
- These rules are formulated as a synchronous grammar

$S \rightarrow \langle X, X \rangle$	glue rule 1
$S \rightarrow \langle S\ X, S\ X \rangle$	glue rule 2
$X \rightarrow \langle \gamma, \alpha, \sim \rangle\ ,\ \gamma, \alpha \in \{X \cup \mathbf{T}\}^+$	hiero rules

Table: Rules contained in the standard hierarchical grammar.

- Gaps have no syntactic meaning
- Greedy automatic extraction from aligned parallel data, with standard constraints[3]
- For search: Hierarchical Cube pruning decoders,...

[2] David Chiang. **A hierarchical phrase-based model for statistical machine translation.** ACL, 2005.

[3] David Chiang. **Hierarchical phrase-based translation.** Computational Linguistics, 2007.

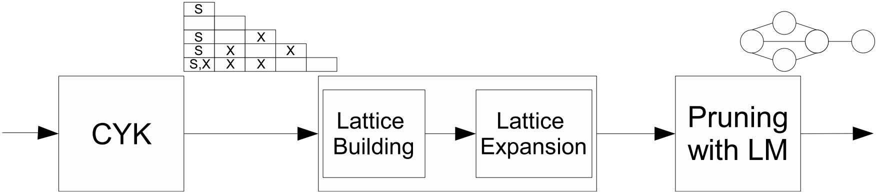
HiFST

- A hierarchical decoder that uses lattices[4] for translation hypotheses.
- Why use lattices?
  - Compactness
  - Easily represented with Weighted Finite-State Transducers (WFSTs).
  - WFSTs: OpenFST, available at openfst.org[5]
  - Weights typically represented as costs under Tropical Semiring $\{\mathcal{R}, \oplus = min, \otimes = +, \overline{0} = \infty, \overline{1} = 0\}$
  - Standard WFST Operations templated over semiring (minimize, compose, prune shortestpath, ...) handle efficiently weights
  - Key for success: RTNs
- WFSTs: succesfully applied to several NLP tasks

[4] Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. **Hierarchical phrase-based translation with weighted finite state transducers.** NAACL 2009.

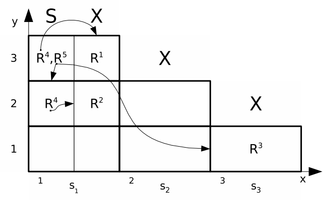
[5] Allauzen, Cyril, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. **OpenFst: A general and efficient weighted finite-state transducer library.** CIAA,2007.

HiFST - General Framework



- CYK algorithm: source side
  - Given a sentence $s_1 ... s_J$, and a synchronous grammar, find all derivations with root in cell $(S, 1, J)$
- Lattices $\mathcal{L}(N, x, y)$ are built for each cell following back-pointers of the grid
  - Objective is the expanded lattice $\mathcal{L}(S, 1, J)$, at the top of the grid
- Apply language model to $\mathcal{L}(S, 1, J)$ and prune

Lattice Building

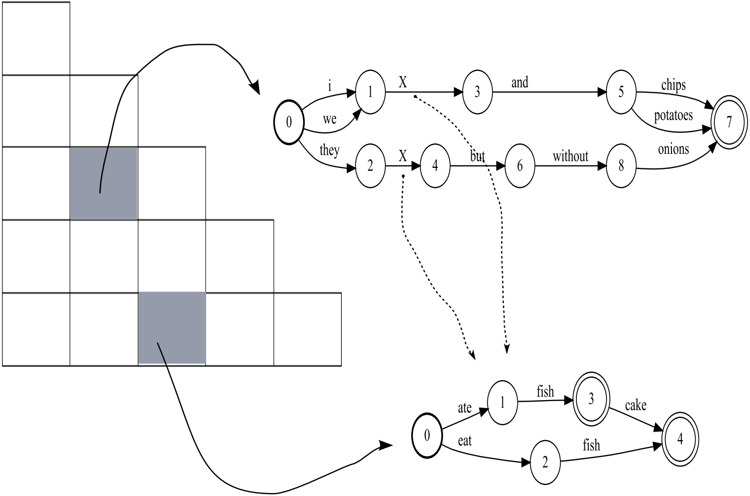



```
function buildFst(N,x,y)
    if L(N,x,y) exists, return L(N,x,y)
    for each rule applied in cell (N,x,y),
      for each element in rule
        if element is a word, create A(element)
          else buildFst(backpointers(element))
      Create rule lattice by catenation of element lattices
    Create cell lattice L(N,x,y) by unioning rule lattices
    Reduce L(N,x,y) with FST operations and return
```

- Recursive algorithm with memoization – traverses grid and returns RTN for the topmost cell. Lower level lattices also stored[6].

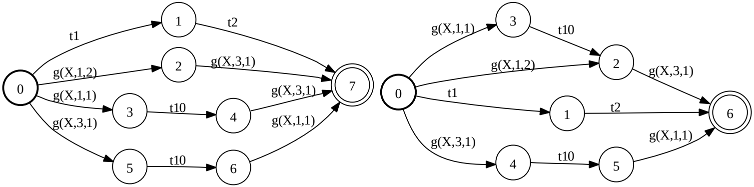
[6] Adrià de Gispert and Gonzalo Iglesias and Graeme Blackwood and Eduardo R. Banga and William Byrne. **Hierarchical Phrase-based Translation with Weighted Finite State Transducers and Shallow-N Grammars.** Computational Linguistics,2010.

Lattice Expansion I



- Our lattices are a mixture of words and pointers to lower level lattices (RTNs)
- Topmost cell lattice $\mathcal{L}(S, 1, J)$ is expanded.

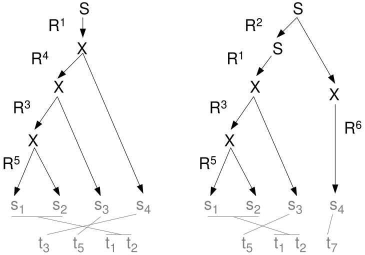
Lattice Expansion II



- Usual operations (rmepsilon, determinize, minimize, etc) work over RTNs (and keep them small) !

Shallow-N grammars I

R^1: $S \rightarrow \langle X , X \rangle$
R^2: $S \rightarrow \langle S\ X , S\ X \rangle$
R^3: $X \rightarrow \langle X\ \mathrm{s}_3 , \mathrm{t}_5\ X \rangle$
R^4: $X \rightarrow \langle X\ \mathrm{s}_4 , \mathrm{t}_3\ X \rangle$
R^5: $X \rightarrow \langle \mathrm{s}_1\ \mathrm{s}_2 , \mathrm{t}_1\ \mathrm{t}_2 \rangle$
R^6: $X \rightarrow \langle \mathrm{s}_4 , \mathrm{t}_7 \rangle$

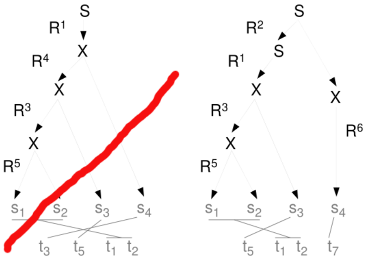


- For certain language pairs, this rule nesting might be unnecessary

Shallow-N grammars II

- Allowing only one level of hierarchical rule nesting is trivial:

R^1: $S \rightarrow \langle X , X \rangle$
R^2: $S \rightarrow \langle S\ X , S\ X \rangle$
R^3: $X_1 \rightarrow \langle X_0$ s$_3$,t$_5$ $X_0 \rangle$
R^4: $X_1 \rightarrow \langle X_0$ s$_4$,t$_3$ $X_0 \rangle$
R^5: $X_0 \rightarrow \langle$s$_1$ s$_2$,t$_1$ t$_2\rangle$
R^6: $X_0 \rightarrow \langle$s$_4$,t$_7\rangle$



- Easily extended to any N levels: Shallow-N grammars.
- Limiting rule nesting to a fixed threshold is a kind of derivation filtering

Performance

- Lattice output has benefits for lattice rescoring and system combination:
  - Large Language Model rescoring
  - Lattice MBR for rescoring and system combination[7] [8]
- Translation tasks between close languages do not require complex rule nesting – Shallow-1 grammars reach similar state-of-the-art performance with much faster decoding times
  - Arabic-to-English[9]
  - Spanish-to-English[10]

[7] S. Kumar and W. Byrne. *Minimum Bayes-risk decoding for statistical machine translation.* NAACL 2004.

[8] R. Tromble, S. Kumar, F. Och, and W. Macherey. *Lattice Minimum Bayes-Risk decoding for statistical machine translation.* EMNLP 2008.

[9] See EACL 2009 paper.

[10] Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. **The HiFST system for the Europarl Spanish-to-English task.** SEPLN, 2009.

Workshop in Machine Translation 2010

- Very competitive translation systems using shallow-1 grammars[11]

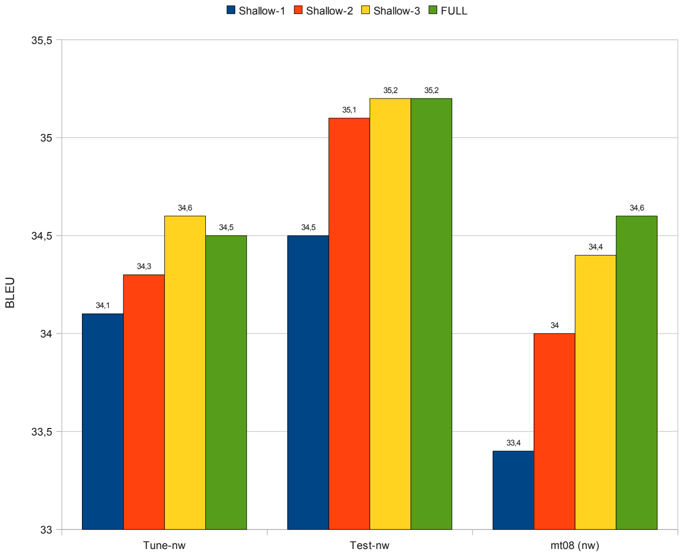
	news-test2008	newstest2009	newstest2010
SP-EN	25.4	27.0	30.5
EN-SP	24.7	25.5	29.1
FR-EN	25.6	29.3	29.6
EN-FR	24.2	26.1	28.2

Table: WMT10 HiFST+LMBR Translation Systems

[11] J. Pino, G. Iglesias, A. de Gispert, G. Blackwood, J. Brunning and W. Byrne. **The CUED HiFST System for the WMT10 Translation Shared Task.** WMT, 2010.

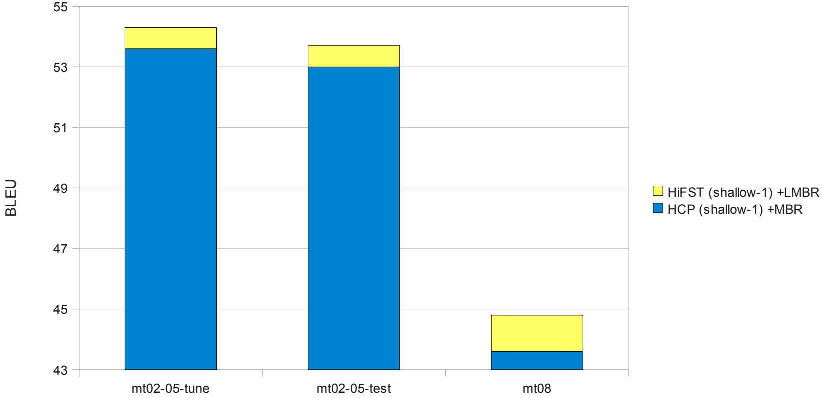
Shallow-N for ZHEN

- Chinese-to-English task: word reordering requirements are harsh.
- Shallow-3 almost bridges the gap at faster decoding times



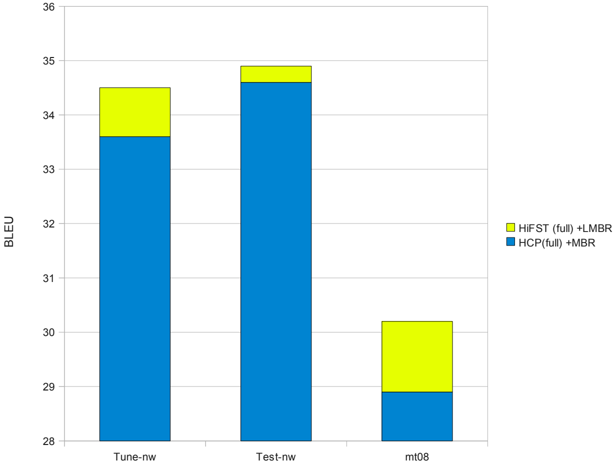
HiFST versus Cube Pruning I

- HiFST with cube pruning decoder for Arabic-to-English translation task (shallow-1). No search errors + lattices = increased performance.



HiFST vs Cube Pruning II

- HiFST with cube pruning decoder for Chinese-to-English translation task (full grammar). Both require local pruning. Still, lattice rescoring methods yield increased performance.



ZHEN task vs. local pruning

- We still have local pruning on ZHEN translation
- Full grammars define search spaces far too big
- Is it possible to avoid local pruning for ZHEN translation tasks?
- Yes: See our next paper in EMNLP 2011[12]! Recipe:
  - Push-down automata
  - 1st pass with entropy pruned language model
  - Rescoring with full language model

[12]Gonzalo Iglesias,Cyril Allauzen,William Byrne Adrià de Gispert, and Michael Riley.**Hierarchical Phrase-based Translation Representations.**To appear in EMNLP 2011.

HiFST goes online – FAUST project

- FAUST:Feedback Analysis for User Adaptive Statistical Translation
- Motivation: Current MT systems do not respond to suggestions for improvement. There are diverse technical reasons for this, including:
  - User feedback tends to be very noisy
  - No research published to date makes explicit how statistical translation and language models can be adapted to benefit from feedback provided by web users
  - No mechanisms exist to identify user feedback of value and inmediately change behaviour of SMT systems in order to avoid the problem
  - Current SMT systems and research efforts are aimed at sophisticated users - translation professionals, intelligence analysts, etc. These users develop an understanding of how to work around their system weaknesses
  - Casual users are tend to be frustrated by a general lack of fluency

HiFST goes online – FAUST project

- HiFST will be available very soon. Check **www.reverso.net**!

Conclusions

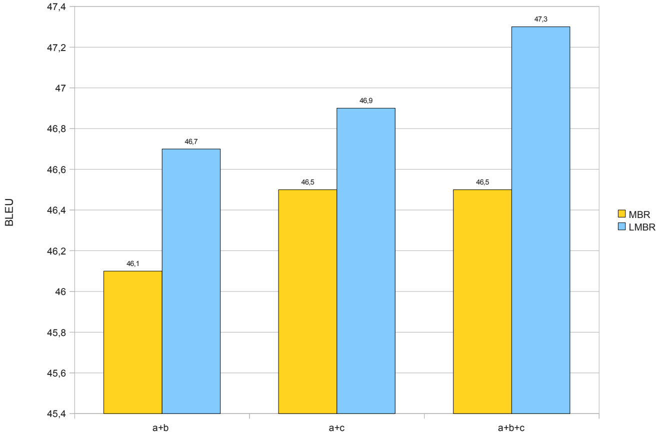
- We described HiFST, a hierarchical decoder based on WFSTs
  - Easy to implement, as complexity is hidden by OpenFST library
  - RTNs effectively reduce complexity during lattice construction
- For ZHEN: Shallow-3 almost bridges the gap with Hiero Full
- ZHEN: bigger search spaces – we still need local pruning
- For languages not in need of strong word reorderings, shallow-1 grammars generally enough
- AREN and WMT10: no search errors: exact decoding

Thank you!

Questions?

System Combination of HiFST systems

- Arabic-to-English NIST09 MT08.
- Three HiFST systems over same Arabic sources with different tokenizations (MADA, SAHKR)
- LMBR combines word lattices and searches for hypotheses with highest similarity to the rest of the lattice



Marginalization over Translation Derivations

- Pruned lattices mapped to log semiring – determinization leads to improved performance
- Improvements do not carry through after LMBR step