Investigations in Exact Inference for Hierarchical Translation

Wilker Aziz[1] Marc Dymetman[2] Sriram Venkatapathy[2]

[1]University of Wolverhampton
Wolverhampton, UK
w.aziz@wlv.ac.uk

[2]Xerox Research Centre Europe
Grenoble, France
{*first.last*}@xrce.xerox.com

August 9, 2013

Table of Contents I

Optimisation in hierarchical translation

Hierarchical phrase-based translation

- SCFG compactly encodes the translation equivalences
- incorporate the language model requires intersecting a wFSA
  - while this is guaranteed polynomial in time and space
  - it is **prohibitive** even for low order LMs
  - "requires" approximation [Chiang, 2007]

Proposal

Avoid performing the full intersection, but **without** losing exactness

1. start from an optimistic unigram LM
2. incorporate higher-order *n*-grams on the basis of evidence of the need to do so

Exact inference over a tractable proxy representation of the target distribution (dynamic programming)

Using a technique that is also directly applicable to **sampling**

Sampling

During **decoding**, when a single output is required, **optimisation** is a natural choice

However,

- Minimum Bayes Risk decoding is based on samples
- samples are also useful for exploring different modes of the distribution

During **learning**, **samples** are necessary for training the parameters

- however, often n-best lists are used as a proxy
  e.g. MERT, minimum risk training

[Blunsom and Osborne, 2008, Arun et al., 2009]

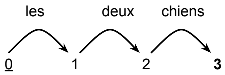
Exact optimisation and sampling with OS*

A unified view on optimisation and sampling

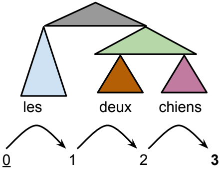
- A cross between Adaptive Rejection Sampling and A* optimisation
- An exact alternative to the usual approximate MCMC sampling techniques (e.g. Gibbs)

OS* [Dymetman et al., 2012]

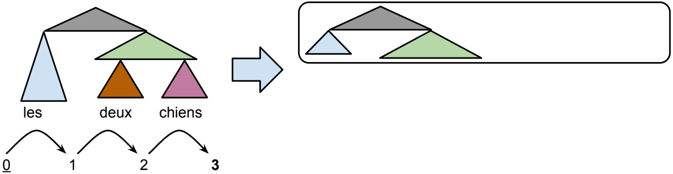
Revisiting hierarchical SMT in 30 seconds



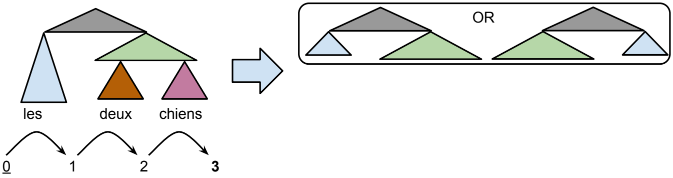
Revisiting hierarchical SMT in 30 seconds



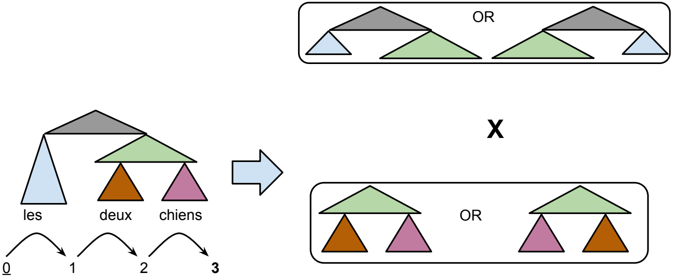
Revisiting hierarchical SMT in 30 seconds



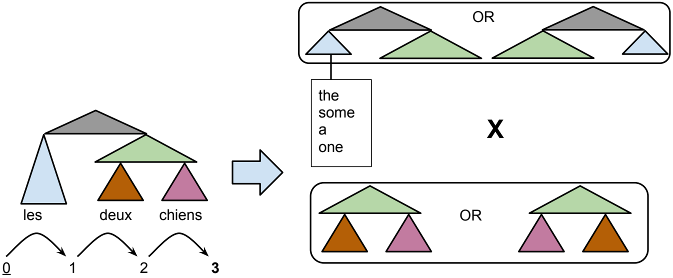
Revisiting hierarchical SMT in 30 seconds



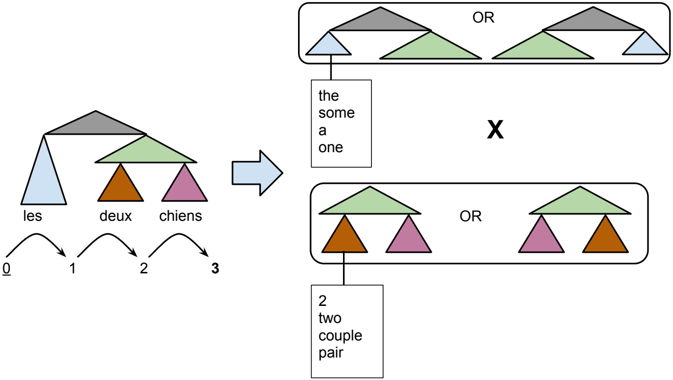
Revisiting hierarchical SMT in 30 seconds



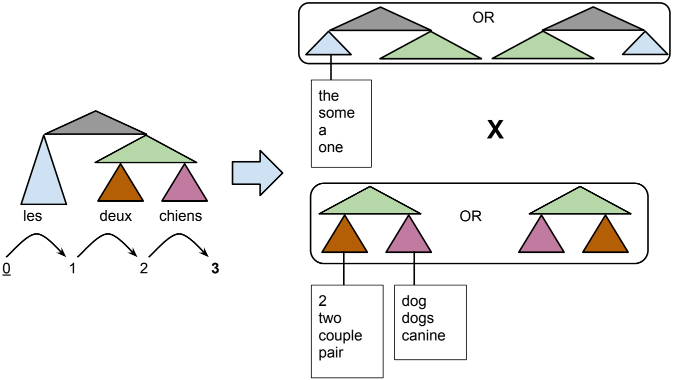
Revisiting hierarchical SMT in 30 seconds



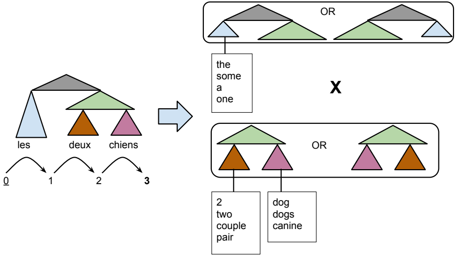
Revisiting hierarchical SMT in 30 seconds



Revisiting hierarchical SMT in 30 seconds



Revisiting hierarchical SMT in 30 seconds

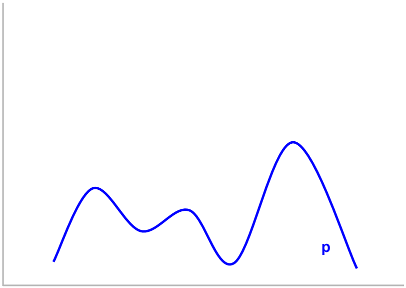


- translation hypergraph without the LM: $G(f)$
  wCFG
- language model: A
  wFSA

$G(f) \cap A$ **intractable** space of weighted translations of the input [Dyer, 2010]

- defines an unnormalised probability distribution over target derivations
- number of rules in the intersected grammar grows exponentially with the order of A

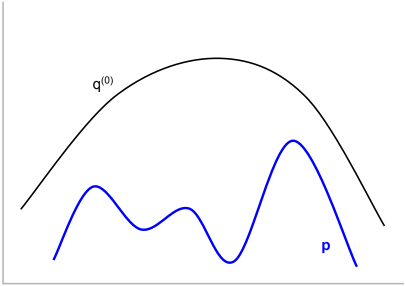
OS* for hierarchical SMT



$p = G(f) \cap A$

intractable $\rightarrow$ ~~**dynamic programming**~~

OS* for hierarchical SMT



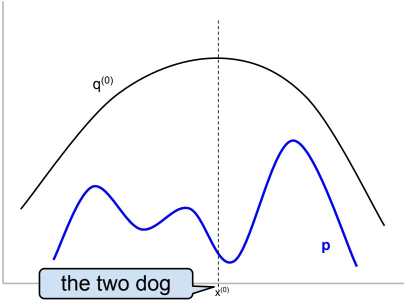
Simpler and optimistic proposal

$$q^{(0)} = G(f) \cap A^{(0)}$$

tractable $\rightarrow$ **dynamic programming**

- $A^{(0)}$ is an optimistic unigram version of the full LM
- Progress by lowering the upper-bound based on observed samples
- Efficient "Earley intersection" [Dyer, 2010]

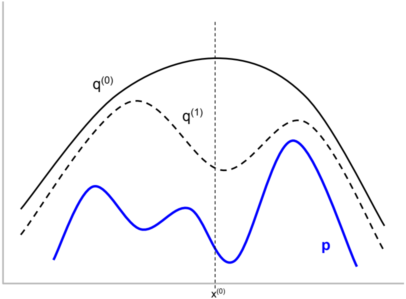
OS* for hierarchical SMT



Sample from $q^{(0)}$

- Accept sample with probability $r = p(x)/q^{(0)}(x)$
- Poor acceptance rate
- Rejected samples are used to refine the proxy

OS* for hierarchical SMT



Obtain a better proxy by intersecting with a small "refinement" automaton $A^{(1)}$

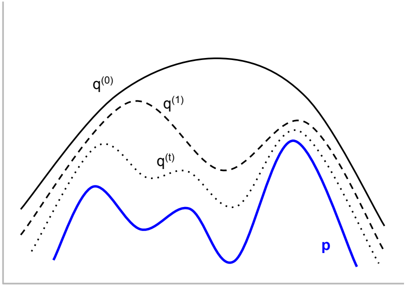
$$q^{(1)} = q^{(0)} \cap A^{(1)}$$

- $A^{(1)}$ accounts for a more precise context

$$\mathrm{w(two)w(dog)} \rightarrow \mathrm{w(two)w(dog|two)}$$

- $q^{(1)}$ is only slightly more complex than $q^{(0)}$ thus dynamic programming remains feasible
- leads to a better acceptance rate

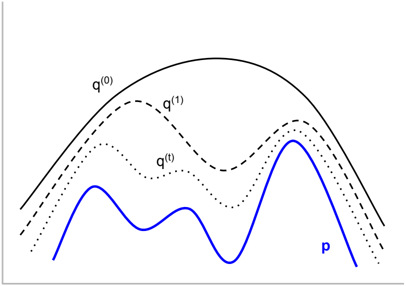
OS* for hierarchical SMT



Repeat the process (sample + refine) until:

- Sampling: a pre-defined acceptance rate is reached

OS* for hierarchical SMT



Repeat the process (sample + refine) until:

- Sampling: a pre-defined acceptance rate is reached
- Optimisation: maximum from $q^{(t)}$ is sufficiently close to p

An upper-bound on the LM distribution

Maximise away the history of an *n*-gram [Carter et al., 2012]

$$w_1(a) \equiv \max_h p_{lm}(a|h)$$

$$w_2(a|a_{-1}) \equiv \max_h p_{lm}(a|h, a_{-1})$$

$$w_3(a|a_{-2}\, a_{-1}) \equiv \max_h p_{lm}(a|h, a_{-2}\, a_{-1})$$

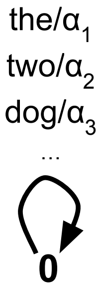
...

Pre-computed

Initial proposal

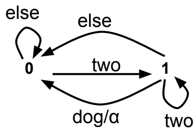
The initial proposal $q^{(0)}$ incorporates only unigrams

- $A^{(0)}$ is a very simple automaton
- $q^{(0)} = G(f) \cap A^{(0)}$ has the same size of $G(f)$



Incremental updates

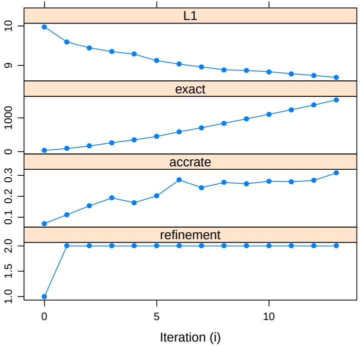
Down-weight occurrences of **dog** *in the context of* **two**



where $\alpha = \frac{w_2(\text{dog}|\text{two})}{w_1(\text{dog})}$

Affects derivations yielding strings that contain occurrences of "two dog"
Each such occurrence is now scaled by α

Illustration (sampling)

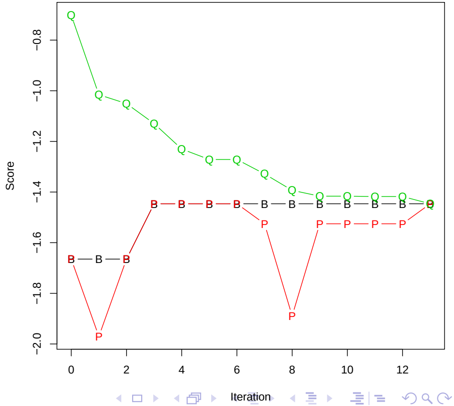


Due to refinements

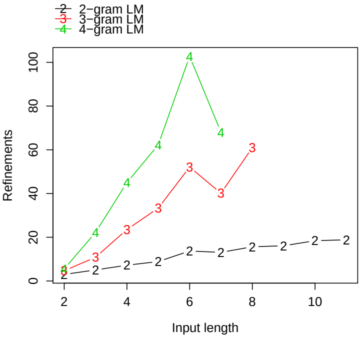
- better acceptance rate

Illustration (optimisation)

i	Rules	Optimum
0	311	<s> one last observation . </s>
1	454	<s> **one** last observation . </s>
2	628	<s> one **last** observation . </s>
3	839	<s> one final **observation** . </s>
4	1212	<s> one **final observation** . </s>
		...
12	3000	<s> **a final** observation . </s>
13	3128	<s> one final observation . </s>



Optimisation

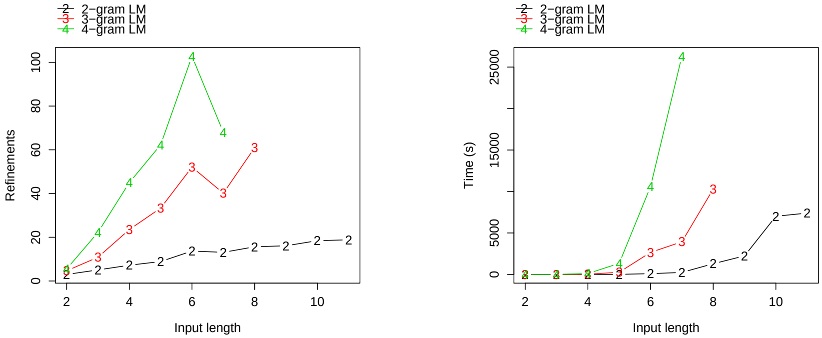


Length	ctxt	count	$\frac{\vert R_f \vert}{\vert R_0 \vert}$
4	1	20.3	74.6 ± 53.9
	2	19.2	
	3	5.4	
5	1	21.9	145.4 ± 162.6
	2	32.9	
	3	7.5	
6	1	**34.7/75**	535.8 ± 480.0
	2	**54.9/2000**	
	3	**13.2**	

4-gram LM

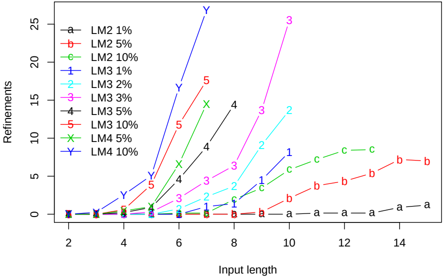
- Needs to account for very few contexts

Optimisation



- Needs to account for very few contexts

Sampling

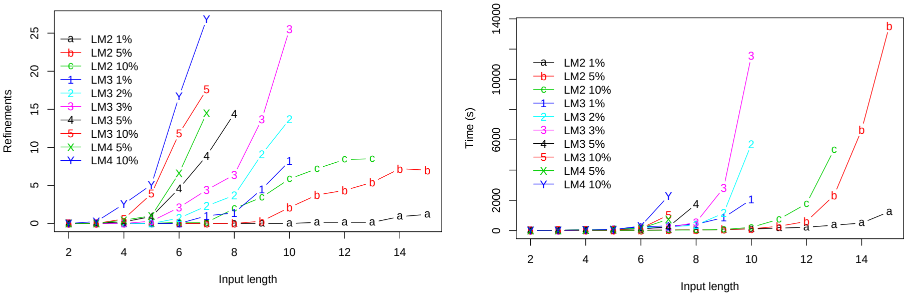


Input	ctxt	count	$\frac{\lvert R_f \rvert}{\lvert R_0 \rvert}$
5	1	1.0	1.9 ± 1.0
6	1	6.3	17.6 ± 13.6
	2	0.3	
7	1	**12.9/90**	93.8 ± 68.9
	2	**1.5/3000**	
	3	**0.1**	

4-gram LM

- Needs to account for very few contexts (mostly lower-order)

Sampling



- Needs to account for very few contexts (mostly lower-order)

Summary

Contributions

- common framework for optimisation and sampling
- exactness
- anytime guarantees: acceptance rate / distance to optimum
- explore only a sub-space of the possible *n*-grams

Challenge: control the time $\rightarrow$ complexity of the intersection

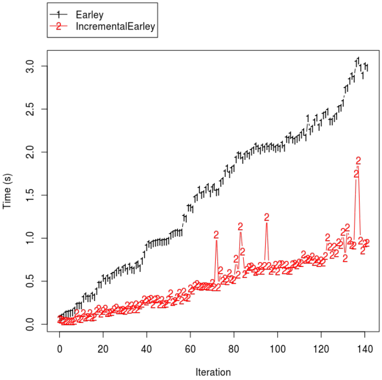
Thanks!

Questions/comments?

Incremental intersection

Reuse chart items compatible with the new automaton

Motivating example



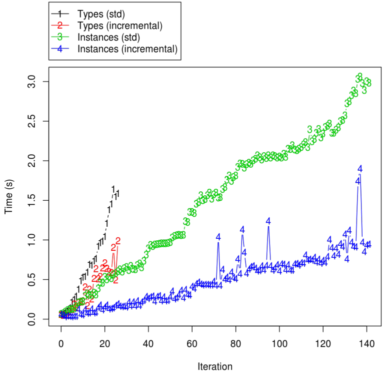
- about a third of the time

Making refinements more local

Distinguish instances of terminals

Motivating example: 47 types [506 instances]

#refinements	#instances	context
2	11	,
5	15	@-@
2	7	central
7	15	centre
1	12	heart
6	57	**in**
5	5	secret
1	8	secretive
4	7	secretly
16	108	**the**
8	41	middle
2	3	midst
5	27	of
1	1	<s>



1-word context:

- types: 314 (62%) instances are affected
- instances: 65 (7%) instances are affected

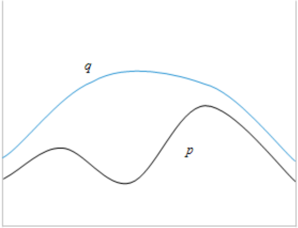
OS*

Exact **O**ptimisation an **S**ampling with Connections to A* (OS*) [Dymetman et al., 2012]

- Coarse-to-fine strategy
- Tractable form of adaptive rejection sampling

OS* (sampling)

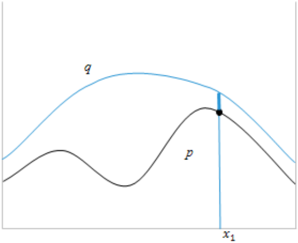
We upper-bound the target distribution p by a simpler proposal q and proceed in a adaptive rejection sampling fashion



- we can optimise/sample from q directly

OS* (sampling)

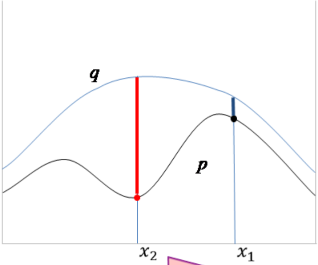
In sampling, we sample from q



- x_1 is accepted with probability $r = p(x_1)/q(x_1)$

OS* (sampling)

In sampling, we sample from q

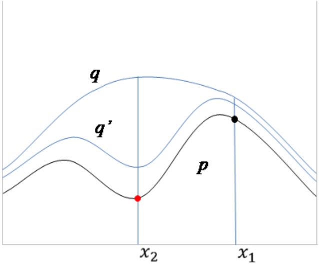


Evidence that we are being too optimistic!
However, not everywhere, rather at around $x = x_2$

- x_1 is accepted with probability $r = p(x_1)/q(x_1)$
- x_2 is rejected

OS* (sampling)

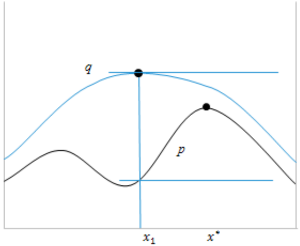
Rejected samples are used to motivate an increase in the complexity of q



- accounts for some underspecified context
- brings the proxy closer to the target
- increases the rate of acceptance

OS* (optimisation)

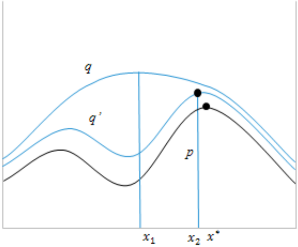
In optimisation, we find the maximum of q



- x_1 is rejected due to low ratio $r = p(x_1)/q(x_1)$

OS* (optimisation)

Rejected maxima are used to motivate an increase in the complexity of q



- accounts for some underspecified context
- brings q's maximum closer to the true maximum

OS* (convergence)

In sampling

- longer contexts are incorporate till a pre-defined acceptance rate is achieved

OS* (convergence)

In sampling

- longer contexts are incorporate till a pre-defined acceptance rate is achieved

In optimisation

- contexts are incorporated while $q(x)$ differs sufficiently from $p(x)$ for $x = \text{argmax}_x q(x)$
  i.e.
  $$r(x) = p(x)/q(x) < \epsilon$$

Chart item combination

In $G' \equiv G(f) \cap A$

Chart item combination

In $G' \equiv G(f) \cap A$

- rules have same length, structure and terminals of those in $G(f)$

Chart item combination

In $G' \equiv G(f) \cap A$

- rules have same length, structure and terminals of those in $G(f)$
- but nonterminals are indexed versions of those in $G(f)$ [Bar-Hillel et al., 1961]
  e.g. (i, N, j) where
  - i and j are states in A
  - N is a nonterminal in the original grammar

Chart item combination

In $G' \equiv G(f) \cap A$

- rules have same length, structure and terminals of those in $G(f)$
- but nonterminals are indexed versions of those in $G(f)$ [Bar-Hillel et al., 1961]
  e.g. (i, N, j) where
  - i and j are states in A
  - N is a nonterminal in the original grammar
- number of states in A grows exponentially with the order of the LM

Related work

- Rush and Collins [2011] address exact decoding in HPB-SMT using Dual Decomposition

Related work

- Rush and Collins [2011] address exact decoding in HPB-SMT using Dual Decomposition
- Blunsom and Osborne [2008] address probabilistic inference (at both decoding and training)

  they sample derivations from a cube pruned space

Related work

- Rush and Collins [2011] address exact decoding in HPB-SMT using Dual Decomposition
- Blunsom and Osborne [2008] address probabilistic inference (at both decoding and training)
  they sample derivations from a cube pruned space
- Arun et al. [2009] introduce a Gibbs sampler for PB-SMT
  MBR training/decoding and approximate "max-translation"

OS* for hierarchical SMT

Produce a sequence of "proposal" grammars which all upper-bound p

OS* for hierarchical SMT

Produce a sequence of "proposal" grammars which all upper-bound p

- $q^{(0)} = G(f) \cap A^{(0)}$

OS* for hierarchical SMT

Produce a sequence of “proposal” grammars which all upper-bound p

- $q^{(0)} = G(f) \cap A^{(0)}$
- $q^{(1)} = q^{(0)} \cap A^{(1)}$

OS* for hierarchical SMT

Produce a sequence of "proposal" grammars which all upper-bound p

- $q^{(0)} = G(f) \cap A^{(0)}$
- $q^{(1)} = q^{(0)} \cap A^{(1)}$
- ...

OS* for hierarchical SMT

Produce a sequence of “proposal” grammars which all upper-bound p

- $q^{(0)} = G(f) \cap A^{(0)}$
- $q^{(1)} = q^{(0)} \cap A^{(1)}$
- $\ldots$
- $q^{(t)} = q^{(t-1)} \cap A^{(t)}$

OS* for hierarchical SMT

Produce a sequence of "proposal" grammars which all upper-bound p

- $q^{(0)} = G(f) \cap A^{(0)}$
- $q^{(1)} = q^{(0)} \cap A^{(1)}$
- $\ldots$
- $q^{(t)} = q^{(t-1)} \cap A^{(t)}$

$A^{(0)}$ is an optimistic unigram version of the full LM

OS* for hierarchical SMT

Produce a sequence of "proposal" grammars which all upper-bound p

- $q^{(0)} = G(f) \cap A^{(0)}$
- $q^{(1)} = q^{(0)} \cap A^{(1)}$
- ...
- $q^{(t)} = q^{(t-1)} \cap A^{(t)}$

$A^{(0)}$ is an optimistic unigram version of the full LM
$A^{(t)}$ is a small automaton that refines $q^{(t-1)}$ relative to some k-gram context not yet made explicit

OS* for hierarchical SMT

Produce a sequence of "proposal" grammars which all upper-bound p

- $q^{(0)} = G(f) \cap A^{(0)}$
- $q^{(1)} = q^{(0)} \cap A^{(1)}$
- $\ldots$
- $q^{(t)} = q^{(t-1)} \cap A^{(t)}$

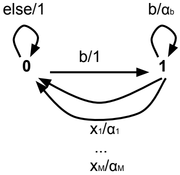
$A^{(0)}$ is an optimistic unigram version of the full LM
$A^{(t)}$ is a small automaton that refines $q^{(t-1)}$ relative to some k-gram context not yet made explicit
Note that for some large M

$$\bigcap_{t=0}^{M} A^{(t)} = A$$

Refining automata

Substring searching construction [Cormen et al., 2001]
Makes a specific **context** explicit

- lower all possible continuations of the history
  e.g. "**b** a", "**b** c", "**b** d"



- note that this does not increase the computational cost of the intersection

Algorithm

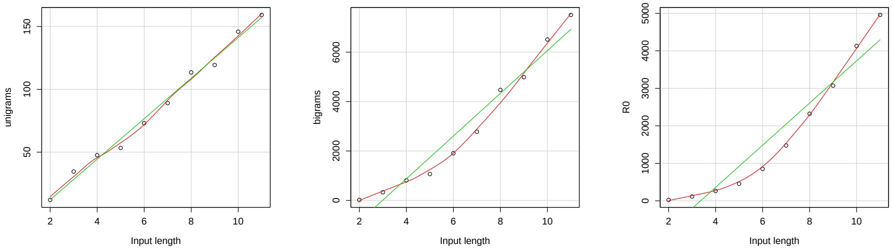
Algorithm 1 OS* for Hierarchical Translation: Optimisation (left) and Sampling (right).

$t \leftarrow 0$
$q^{(0)} \leftarrow G(f) \cap A^{(0)}$
while not an x has been accepted **do**
    Find maximum x in $q^{(t)}$
    $r \leftarrow p(x)/q^{(t)}(x)$
    Accept-or-Reject x according to r
    if Rejected(x) **then**
        define $A^{(t+1)}$ based on x and $q^{(t)}$
        $q^{(t+1)} \leftarrow q^{(t)} \cap A^{(t+1)}$
        $t \leftarrow t+1$
return x

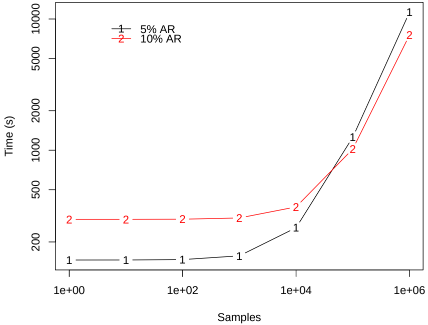
$t \leftarrow 0, AR \leftarrow 0$
$q^{(0)} \leftarrow G(f) \cap A^{(0)}$
while not $AR > threshold$ **do**
    Sample $x \sim q^{(t)}$
    $r \leftarrow p(x)/q^{(t)}(x)$
    Accept-or-Reject x according to r
    if Rejected(x) **then**
        define $A^{(t+1)}$ based on x and $q^{(t)}$
        $q^{(t+1)} \leftarrow q^{(t)} \cap A^{(t+1)}$
        $t \leftarrow t+1$
return already accepted x's along with $q^{(t)}$

Experiment

Small scale experiment: short sentences
Properties of G

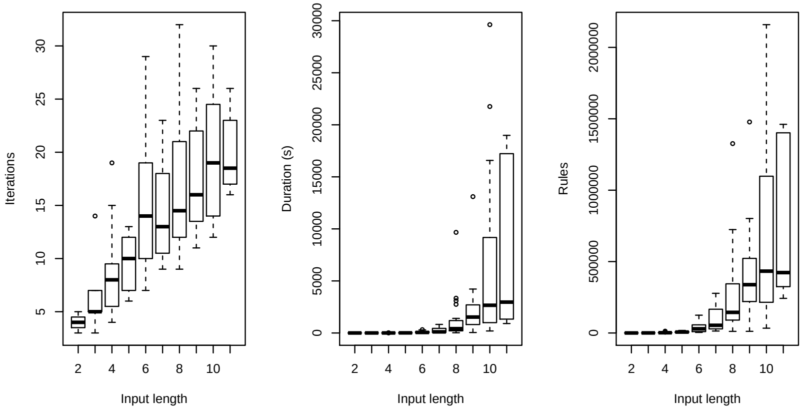


Sampling performance: 4-gram LM

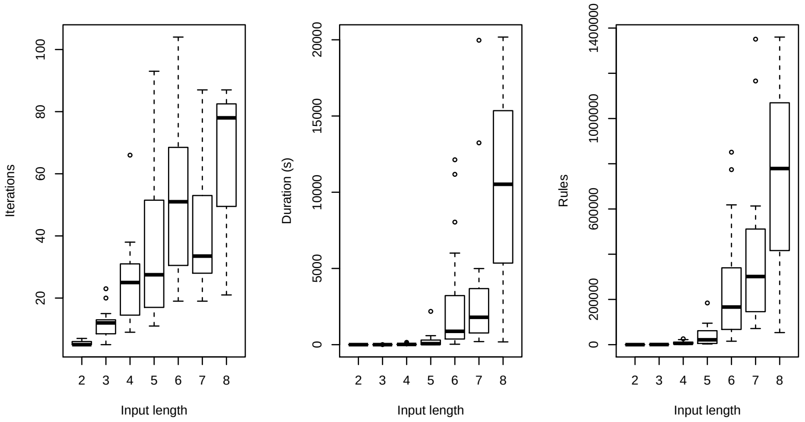


- 20 sentences of length 6
- time to draw 1M samples
- including the time to produce the sampler

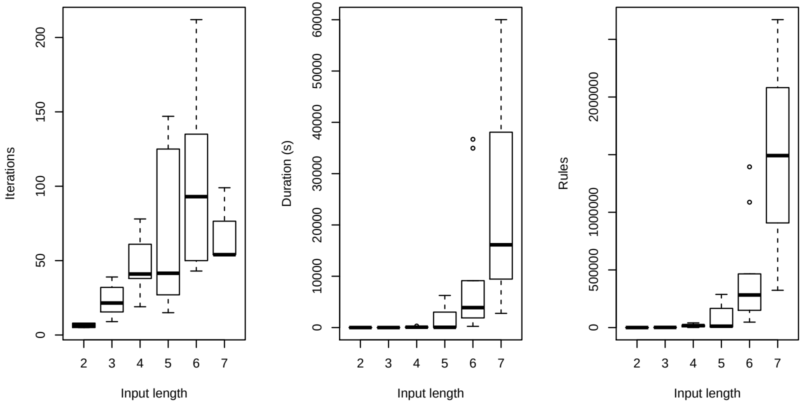
Optimisation (closer look): 2-gram LM



Optimisation (closer look): 3-gram LM



Optimisation (closer look): 4-gram LM



References I

Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez, and Philipp Koehn. Monte carlo inference and maximization for phrase-based translation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 102–110, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-29-9. URL `http://dl.acm.org/citation.cfm?id=1596374.1596394`.

Yehoshua Bar-Hillel, Micha A. Perles, and Eli Shamir. On formal properties of simple phrase structure grammars. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, (14):143–172, 1961.

References II

Phil Blunsom and Miles Osborne. Probabilistic inference for machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 215–223, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613746.

Simon Carter, Marc Dymetman, and Guillaume Bouchard. Exact Sampling and Decoding in High-Order Hidden Markov Models. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1125–1134, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

References III

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270, 2005.

David Chiang. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33:201–228, 2007. URL `http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.2.201`.

Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001. ISBN 0070131511.

References IV

Christopher Dyer. *A Formal Model of Ambiguity and its Applications in Machine Translation*. PhD thesis, University of Maryland, 2010.

Marc Dymetman, Guillaume Bouchard, and Simon Carter. Optimization and sampling for nlp from a unified viewpoint. In *Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology*, pages 79–94, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL http://www.aclweb.org/anthology/W12-6106.

References V

Alexander M. Rush and Michael Collins. Exact decoding of syntactic translation models through lagrangian relaxation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 72–82, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL `http://dl.acm.org/citation.cfm?id=2002472.2002482`.

References VI

Dekai Wu. Stochastic inversion transduction grammars with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of the 14th international joint conference on Artificial intelligence*, volume 2, pages 1328–1335, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.