



Automatic Evaluation of Generation and Parsing for MT with Automatically Induced Transfer Rules

Yvette Graham, Deirdre Hogan and Josef van Genabith
National Centre for Language Technology
School of Computing
Dublin City University

Overview



1. Transfer-based MT
2. MT Evaluation
3. Conventional Parsing/Generation Evaluation Methods
4. New Parsing/Generation Evaluation Method
5. Experiments
 - Hand-crafted LFG technologies
 - Automatically-induced LFG technologies
6. Conclusions



1. Transfer-based Machine Translation

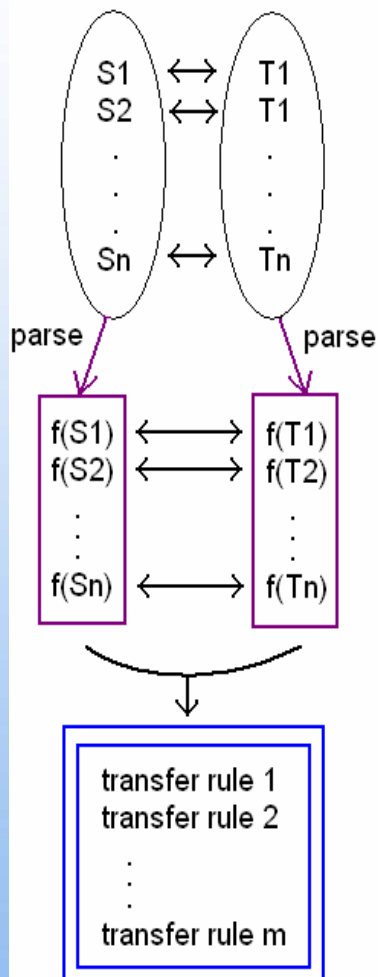
DCU

Transfer-based MT



An oft-cited future application of parsing and generation.

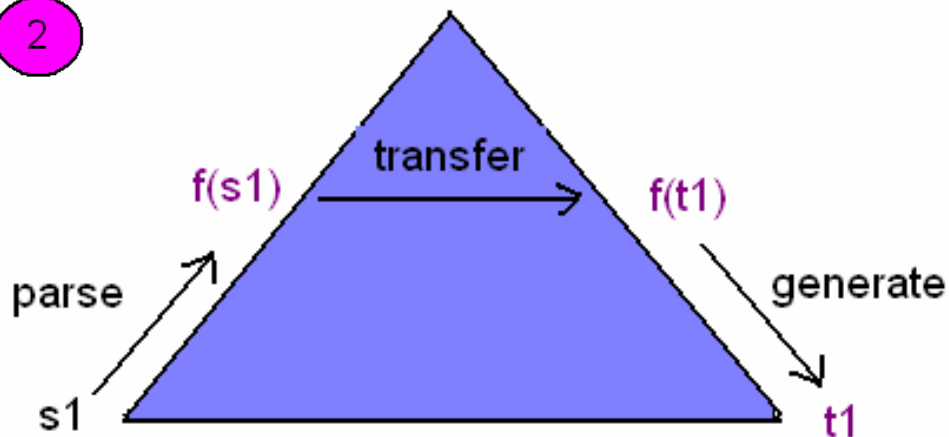
1



Relies heavily on the employed

- parsing technologies
- generation technologies
- how well these technologies work together

2





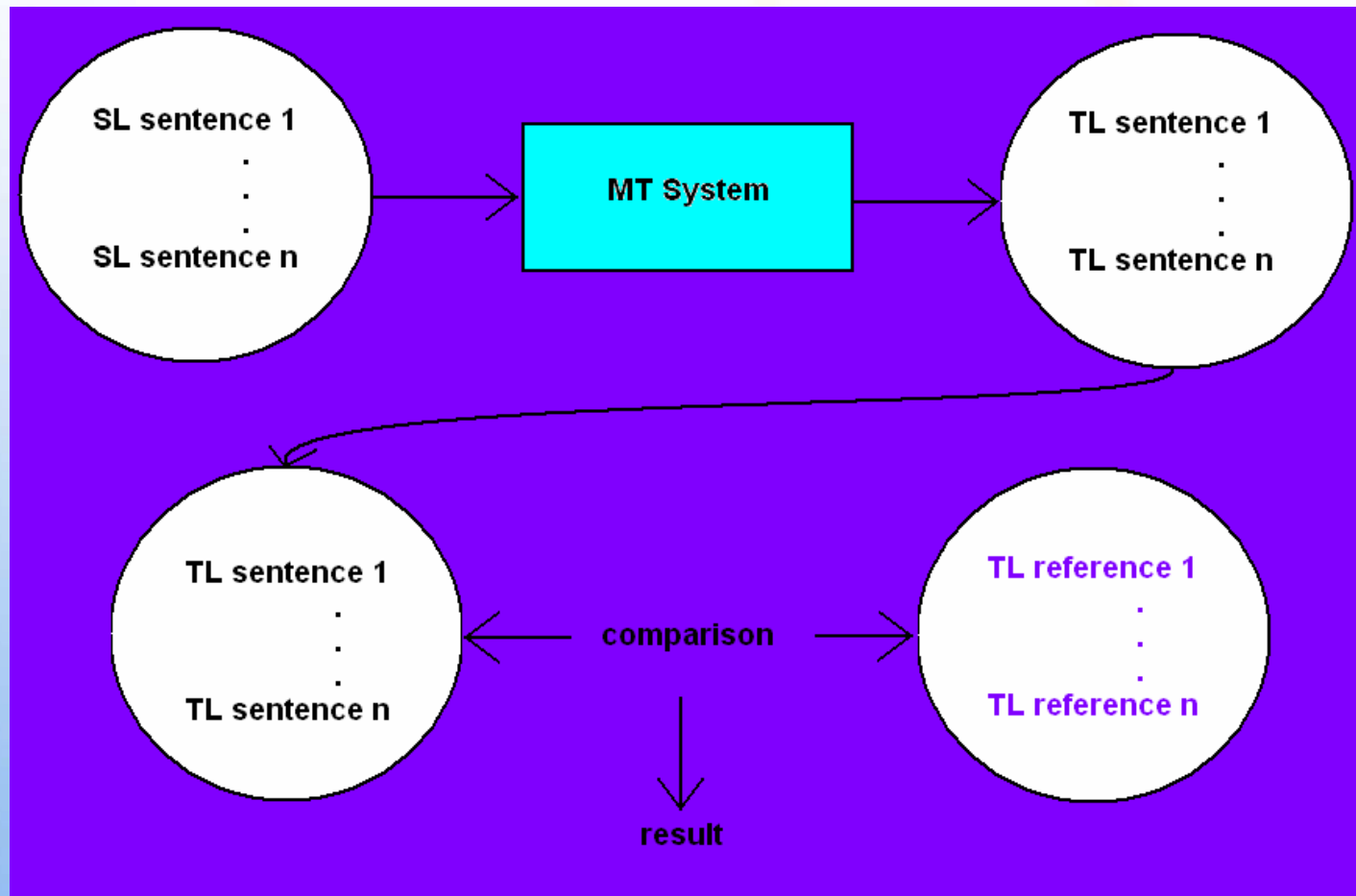
2. Machine Translation Evaluation

DCU

MT Evaluation in General



- Conventional methods evaluate MT system giving an **overall** result for the system on a given test set



Transfer-based MT Evaluation



To understand results more fully

- Need to evaluate parsing & generation technologies in **isolation** from MT system
- But still in a **relevant setting** for transfer-based MT system

Previous work

- Furuse & Iida 1992, Meyers et al. 1998, Menezes & Richardson 2001
- Riezler & Maxwell 2006
 - Training data 80% full parses, 20% fragment parses
 - Test data “only 44% were in coverage of both parsing and generation technologies”



3. Conventional Parsing and Generation

Evaluation Methods and Transfer-based MT

DCU

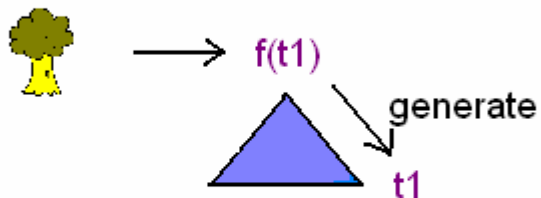
Conventional Evaluation Methods



Why not use Conventional Methods of Evaluation?

Conventional Generation Evaluation:

1:

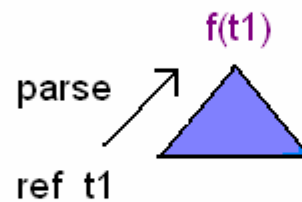


2:

$ref_t1 \leftarrow compare \rightarrow t1$

Conventional Parsing Evaluation:

1:



2:

$f(t1) \leftarrow compare \rightarrow gold\ standard\ f(t1)$

Conventional Parsing Evaluation

National Centre for Language Technology



Why not use Conventional Methods of **Parsing** Evaluation methods?

- **Gold standard** abstract structures of test data required
- Difficult to apply to **new test set**
- Results from different test set, probably **different language domain**
- Does not evaluate how well the **parsing and generation** components work **together**



4. New Evaluation Method for Parsing and Generation

DCU

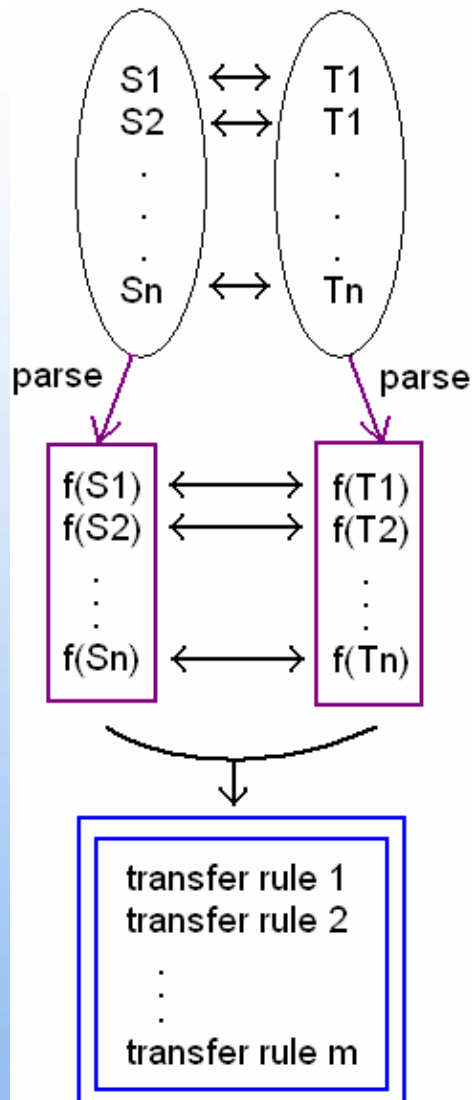
New Evaluation Method



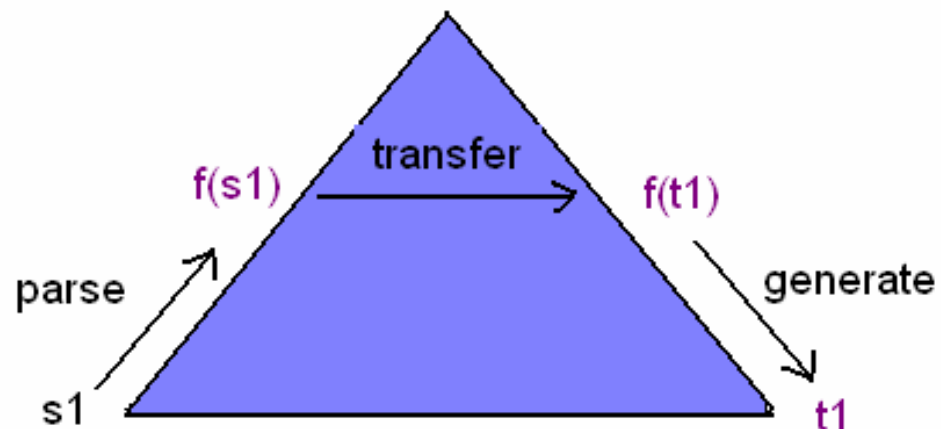
New Method of Parsing / Generation Evaluation

- Evaluate **parsing/generation technologies** that are cited as having **transfer-based MT** as a future application
- Include the **dependence** of the **generation technologies on parsing** technologies
- Easy evaluation on **real MT test data**
- Estimate the **upper bound** imposed on the MT system by these technologies

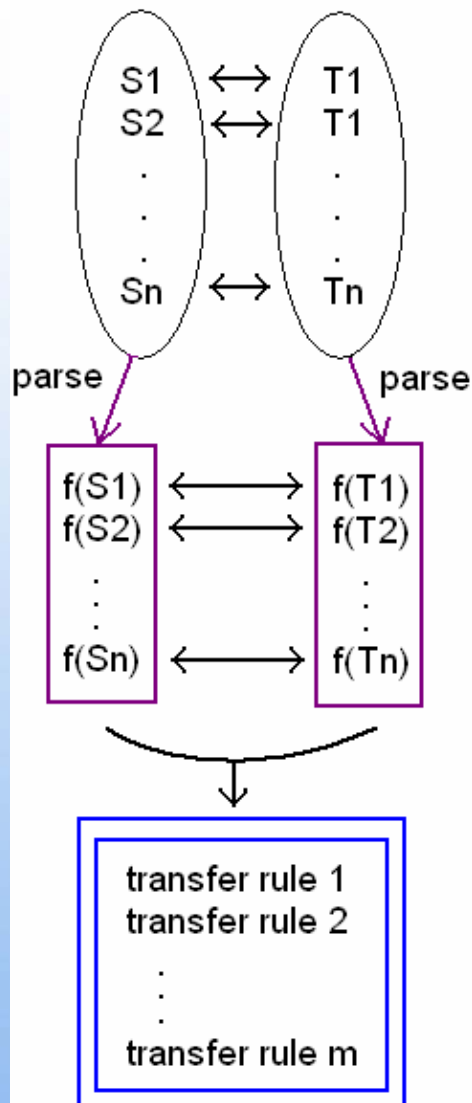
New Evaluation Method



If we assume every part of the transfer MT system is perfect except for the employed TL parsing and generation technologies, What would $f(t_1)$, i.e. the input to the generator be?

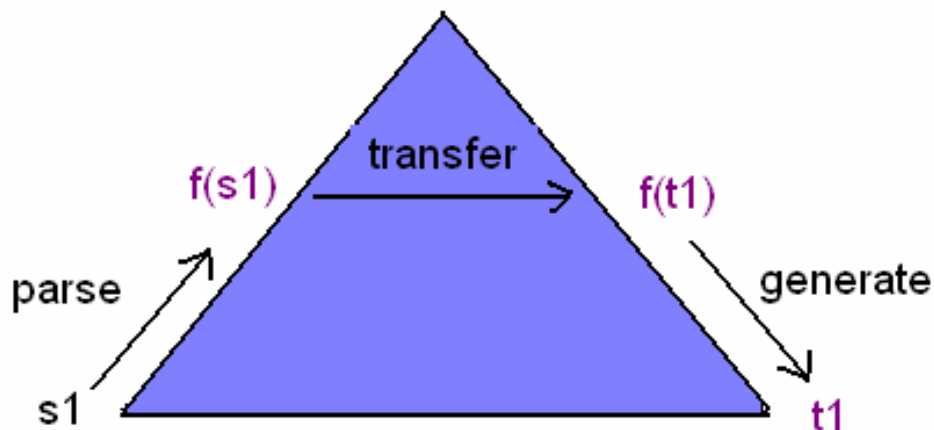


New Evaluation Method



If we assume every part of the transfer MT system is perfect except for the employed TL parsing and generation technologies, What would $f(t_1)$, i.e. the input to the generator be?

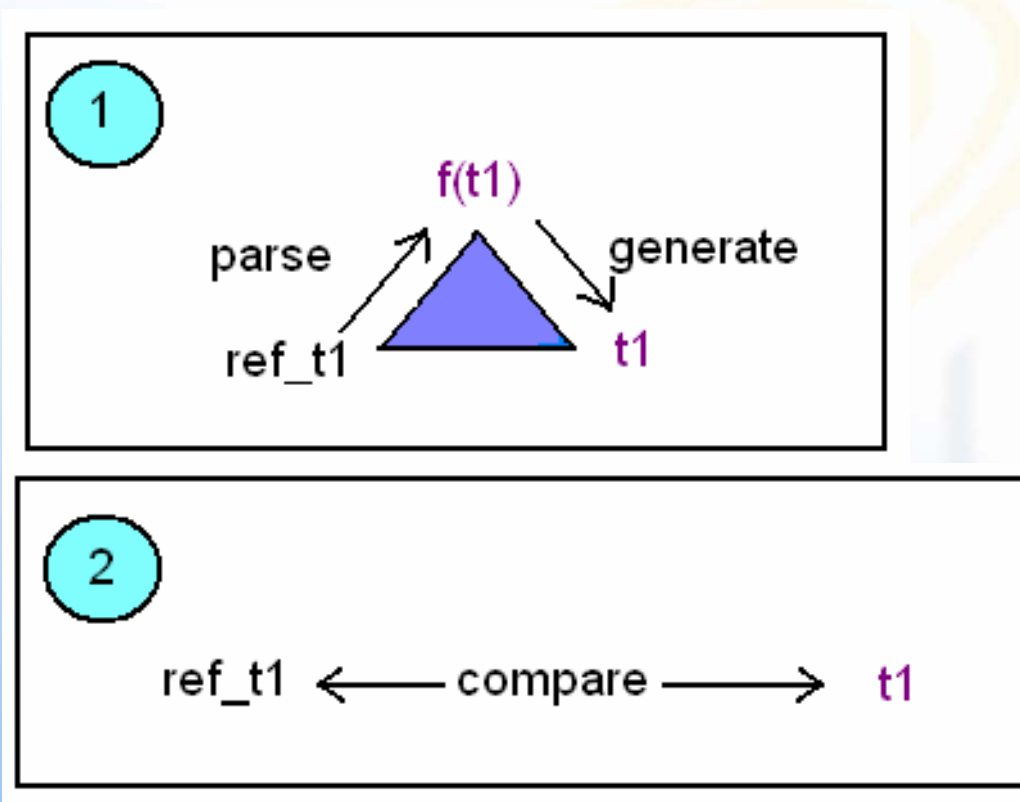
Answer: Parsed ref_t1



New Evaluation Method



To estimate the upper bound imposed by parsing/generation components of MT system:



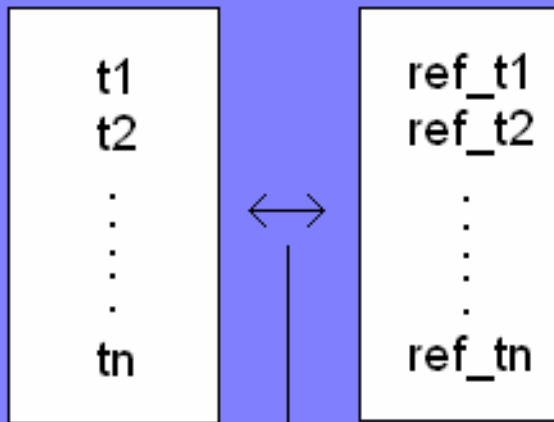


5. Experiment Results





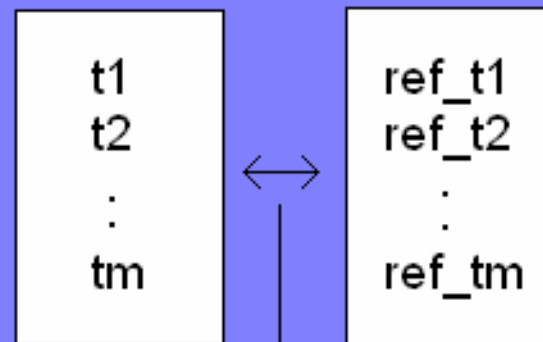
Entire Test Set Results:



score

Example: NIST=9.1

In-coverage only Results:



score, coverage= m/n

Example: NIST=10.4,
coverage=90%



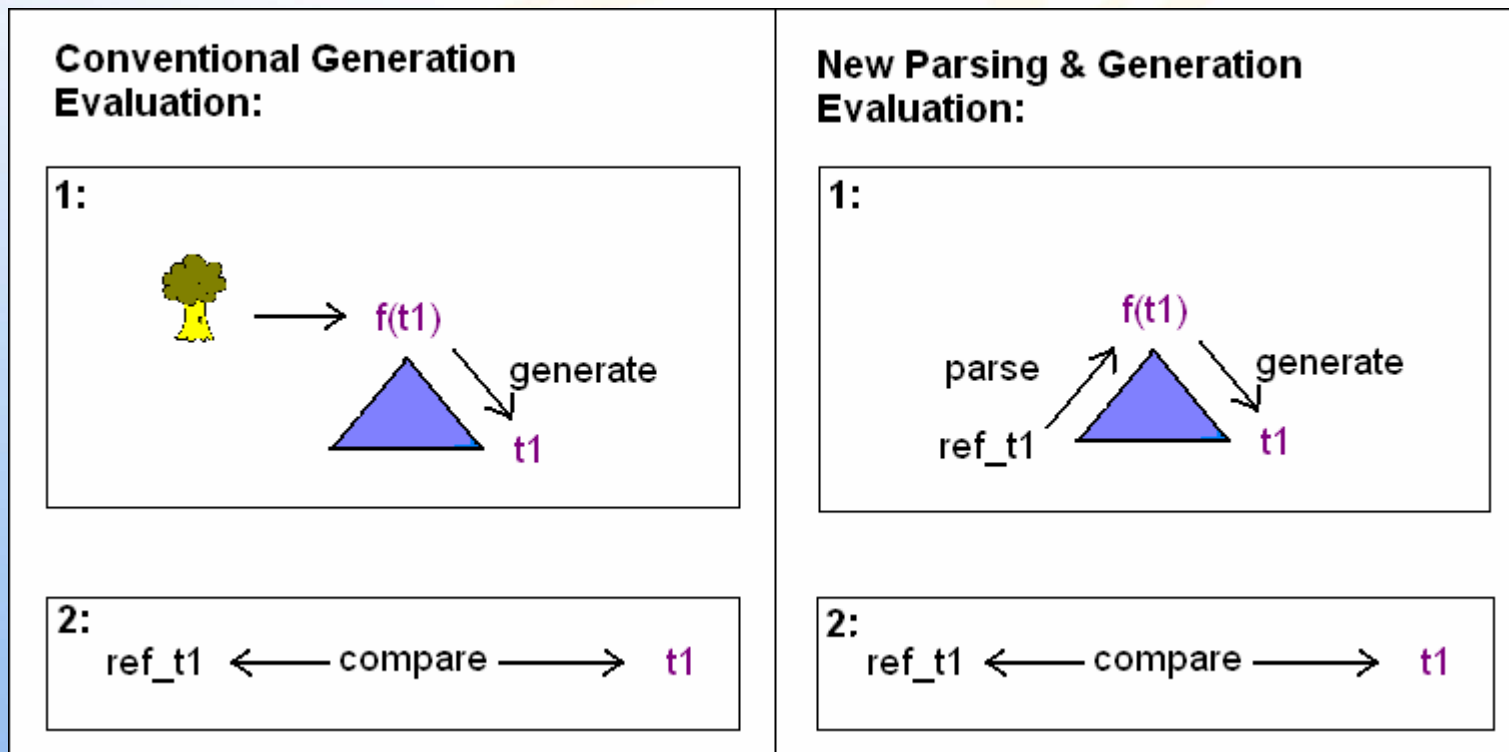
5.1 Experiment 1



Experiment 1: Aim



compare $\left\{ \begin{array}{l} \text{results of evaluating using new method} \\ \text{results of evaluating using conventional generation method} \end{array} \right.$





Evaluated parsing and generation of technologies

- parsing: Cahill et al. (2004)
- generation: Cahill et al. (2006), Hogan et al. (2007)
- automatically induced
- LFG
- English
- trained on WSJ Sections 2-21

Test Data:

- WSJ Section 23

Generator Input:

- Generation Evaluation: gold standard parse tree
- New Evaluation method: raw text sentence

Experiment 1: Entire Test set Results



- Cahill et al. (2004), Cahill et al. (2006), Hogan et al. (2007)
- Entire test set results

	Section 23 (2416 sentences)	
	NIST	BLEU
From Gold-standard Trees	13.29	0.6680
From Parser Trees	13.01	0.6511

- Significant difference for NIST and BLEU scores

Note: domain of language is same as training data

Experiment 1: In-coverage only Results



- Cahill et al. (2004), Cahill et al. (2006), Hogan et al. (2007)
- In-coverage only results

	Section 23 (2416 sentences)		
	NIST	BLEU	Coverage
From Gold-standard Trees	13.31	0.6693	99.88%
From Parser Trees	13.02	0.6515	99.96%

- High coverage
- Scores similar to entire testset scores



5.2 Experiment 2

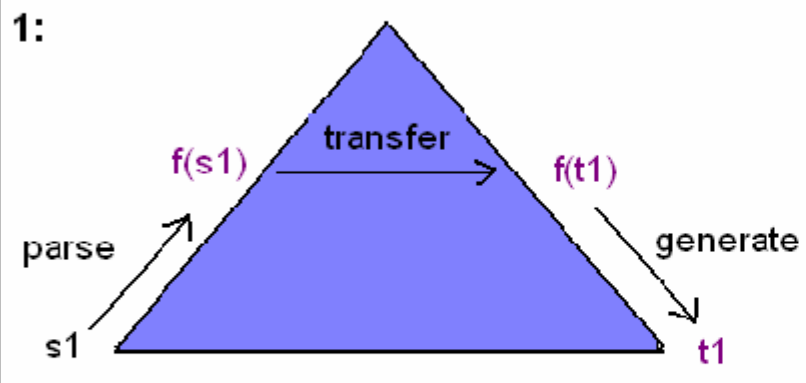


Experiment 2: Aim

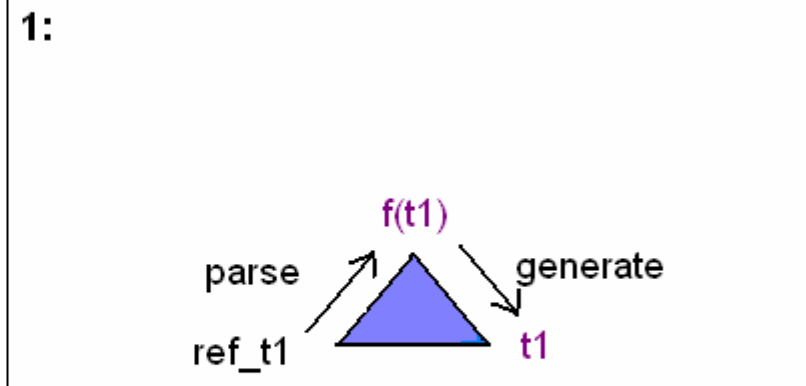


compare $\left\{ \begin{array}{l} \text{results of a transfer-based MT system} \\ \text{TL parser/generator upper bound} \end{array} \right.$

MT System Results:



New Parsing & Generation Evaluation:





MT System Evaluation

- Riezler & Maxwell (2006)
- MT system uses parsing / generation technologies of Riezler et al. (2002)
- Published results: NIST

Employed TL parsing/generation technologies of MT System

- hand-crafted
- LFG f-structures
- English

Test Data:

- English Europarl test set (Koehn et al 2003) sentences length 5-15

Experiment 2: Entire Test set Results



		NIST	BLEU
Europarl (5-15)	MT System (Riezler & Maxwell 2006)	5.62	
	TL Parsing/Generation Upper bound (Riezler et al. 2002)	12.08	0.7431

- High Upper bound → good news for transfer-based MT 😊



5.3 Experiment 3



Experiment 3: Aims A



parsing/generation results of
hand-crafted technologies
using new method on

MT test set 1



MT test set 2



MT test set 3

parsing/generation results of
treebank-induced technologies
using new method on

MT test set 1



MT test set 2



MT test set 3

Experiment 3: Aims B



parsing/generation results of
hand-crafted technologies
using new method on

————— MT test set X



parsing/generation results of
treebank-induced technologies
using new method on

————— MT test set X



Parsing and generation technologies 1

- Riezler et al. (2002)
- hand-crafted
- LFG f-structures
- English

Parsing and generation technologies 2

- Cahill et al. (2004), Cahill et al. (2006), Hogan et al. (2007)
- Treebank-induced
- LFG f-structures
- English



Test Data:

- Europarl test set (Koehn et al 2003) length 5-15 (1755 sentences)
- Europarl test set (Koehn et al 2003) all lengths (500 sentences)
- Homecenter Corpus all lengths (766 sentences)

String Comparison

- BLEU
- NIST

Experiment 3: Entire Test Set Results



	NIST		BLEU	
	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-Induced	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-induced
Europarl (5-15)	12.08	11.72	0.743	0.697
Europarl (all lengths)	6.33	10.24	0.479	0.572
Homecentre	10.75	10.06	0.752	0.6640

- Results vary greatly from one test set to the next for both hand-crafted and induced technologies
 - but more dramatically for hand-crafted technologies

Experiment 3: Entire Test set Results cont.



	NIST		BLEU	
	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-Induced	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-induced
Europarl (5-15)	12.08	11.72	0.743	0.697
Europarl (all lengths)	6.33	10.24	0.479	0.572
Homecentre	10.75	10.06	0.752	0.6640

- Short sentences: hand-crafted better

Experiment 3: Entire Test set Results cont.



	NIST		BLEU	
	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-Induced	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-induced
Europarl (5-15)	12.08	11.72	0.743	0.697
Europarl (all lengths)	6.33	10.24	0.479	0.572
Homecentre	10.75	10.06	0.752	0.6640

- Unrestricted length: automatically-induced technologies better

Experiment 3: Entire Test Set Results cont.



	NIST		BLEU	
	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-Induced	Riezler et al. 2002 Hand-crafted	Cahill et al. 2004 Treebank-induced
Europarl (5-15)	12.08	11.72	0.743	0.697
Europarl (all lengths)	6.33	10.24	0.479	0.572
Homecentre	10.75	10.06	0.752	0.6640

- Homecenter Corpus
 - development data for hand-crafted technologies
- Domain of language: printer manual
 - many imperative sentences
 - not common in WSJ text

Experiment 3: In-coverage only Results



	Cahill et al (2004), Cahill et al. (2006), Hogan et al. (2007) Treebank-induced			Riezler et al. (2002) Hand-crafted		
	NIST	BLEU	Coverage	NIST	BLEU	Coverage
WSJ Section 23	13.02	0.6511	99.96%			
Europarl (5-15)	11.72	0.6968	100%	12.26	0.7800	95%
Europarl (all lengths)	10.24	0.5716	100%	12.1	0.7591	80%
Homecentre	10.06	0.6640	100%	10.81	0.7931	98%

- Effect of **change of domain** on automatically induced resources

WSJ newspaper text → Europarl parliamentary proceedings

Experiment 3: In-coverage only Results



	Cahill et al (2004), Cahill et al. (2006), Hogan et al. (2007) Treebank-induced			Riezler et al. (2002) Hand-crafted		
	NIST	BLEU	Coverage	NIST	BLEU	Coverage
WSJ Section 23	13.02	0.6511	99.96%			
Europarl (5-15)	11.72	0.6968	100%	12.26	0.7800	95%
Europarl (all lengths)	10.24	0.5716	100%	12.1	0.7591	80%
Homecentre	10.06	0.6640	100%	10.81	0.7931	98%

- Treebank-induced technologies → better coverage
- Cannot compare NIST and BLEU scores here → different sentences when coverage not 100%

Conclusions

- Presented a **new method of evaluation** for parsing and generation technologies
- Useful for work that cites transfer-based MT as a **future application**
- Easily applied to any test set – **no gold standard** needed for this evaluation method
- Evaluation Method estimates the **upper bound** imposed by the quality of the TL parsing and generation technologies on a transfer-based MT system
- Provides a **realistic evaluation** in the context of transfer-based MT
- Means simple and quick means of investigating the **viability of a transfer-based MT system** given a particular pair of TL parsing / generation technologies



Thanks!

DCU