BLEUÂTRE: Flattening Syntactic Dependencies for MT Evaluation

Dennis N. Mehay
and
Chris Brew

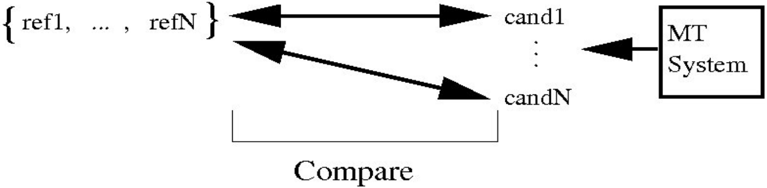
Department of Linguistics
The Ohio State University
{mehay,cbrew}@ling.osu.edu

Theoretical and Methodological Issues in MT (2007)
Skövde, Sweden

Outline

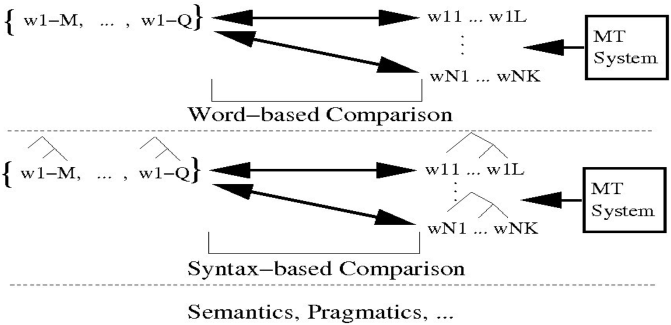
1. Target Language-based MT Evaluation: The Basic Regime
2. A Tour of Other Approaches: Motivating BLEUÂTRE
   - BLEU and NIST: N-gram-based MT Evaluation
   - METEOR
   - Syntax-based Approaches
3. BLEUÂTRE: Flattening and Using Word-word Dependencies
4. Experiments with LDC TIDES Multiple Translation "Chinese"

(Thompson, 1991) Comparing Candidates to References



- Reference (target language) corpus is one-time investment.
- Comparison is consistent and (potentially) fast, cheap, etc.

Ways of Comparing Candidates to References

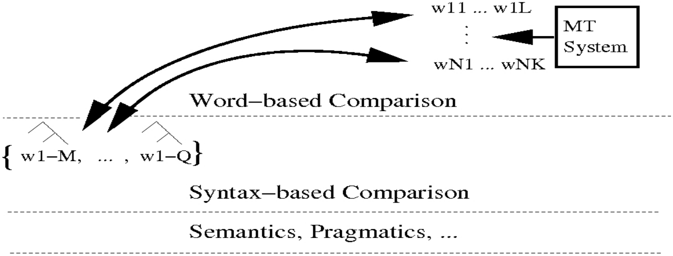


- Word-based is well-represented — (Thompson, 1991; Brew and Thompson, 1994), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), etc.
- Synax-based is gaining traction — (Liu and Gildea, 2005), (Owczarzak et al., 2007).

Simulating Parsing: Combining Syntax- and Word-based Technologies

- Is there a middle ground?
- How do you use parse information from references *without* parsing the candidates?
- Cf. TextRunner (Banko et al., 2007) ⇒ they *simulate parsing* by training word- and POS-fed classifiers to recognise dependencies in strings.
- We want to simlulate parsing in a similar way.

Our Approach: BLEUÂTRE ('Bluish')



- Use syntactic information from reference set.
- "Compile" it down to a form suitable for word-based comparison.
- Motivation: Draw on strengths of word- and syntax-based approaches.
  - Avoid parsing where possible.
  - But only look for *syntactically relevant* word matches.

BLEU and NIST

- Measure translation quality by n-gram overlap with reference(s).
- Typically $1 \leq n \leq 4 \text{ or } 5$
- Strengths:
  - Simple, fast and cheap: only word matching.
  - Portable: only have to port (or develop) tokenisers.
  - Reference set is (virtually) the only investment.
- Shortcomings:
  - Sometimes do not correlate with human judgments (Callison-Burch et al., 2006)
  - Behavior is unreliable in presence of (good and bad) word-order variation.

BLEU and NIST: How to break them.

- Some words can "move around", some cannot. BLEU and NIST do not distinguish the two cases.

Reference(s)	**Candidates**	
Please fill your name in ...	c1: *Fill please your name in* c2: *Please fill in your name* c3: *Please fill your name in* ...	

Figure: (Key: unigram, bigram, trigram and 4-gram match(es).)

BLEU and NIST: How to break them.

- Some words can "move around", some cannot. BLEU and NIST do not distinguish the two cases.

Reference(s)	**Candidates**	
Please fill your name in ...	c1: *Fill please your name in* c2: *Please fill in your name* c3: *Please fill your name in* ...	⇐ perfectly good.

Figure: (Key: unigram, bigram, trigram and 4-gram match(es).)

BLEU and NIST: How to break them.

- Some words can "move around", some cannot. BLEU and NIST do not distinguish the two cases.

Reference(s)	**Candidates**	
Please fill your name in	c1: *Fill please your name in*	⇐ this scores higher
...	c2: *Please fill in your name*	⇐ perfectly good.
	c3: *Please fill your name in*	
	...	

Figure: (Key: unigram, bigram, trigram and 4-gram match(es).)

BLEU and NIST: How to break them.

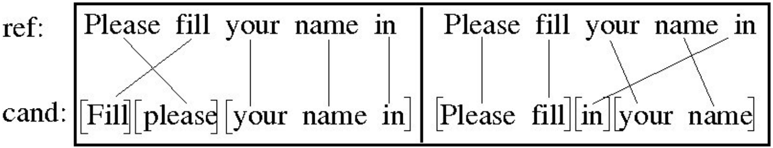
- Some words can "move around", some cannot. BLEU and NIST do not distinguish the two cases.

Reference(s)	**Candidates**	
Please fill your name in ...	c1: *Fill please your name in* c2: *Please fill in your name* c3: *Please fill your name in* ...	⇐ this scores higher ⇐ perfectly good.

Figure: (Key: unigram, bigram, trigram and 4-gram match(es).)

- (Callison-Burch et al., 2006): w.r.t. one reference, can be $> 10^{73}$ permutations of a sentence with same BLEU score (or better).

METEOR: Susceptible to the Same Word-order Pitfalls

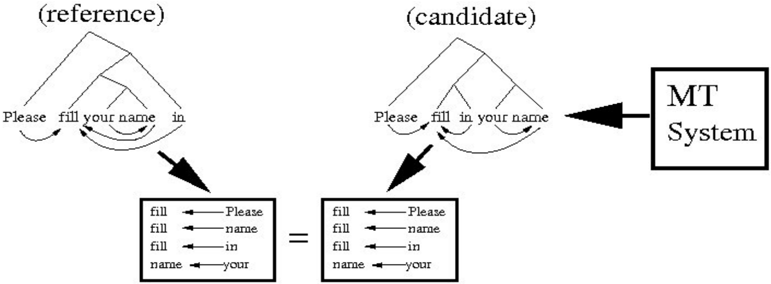


ref: Please fill your name in | Please fill your name in

cand: [Fill][please][your name in] | [Please fill][in][your name]

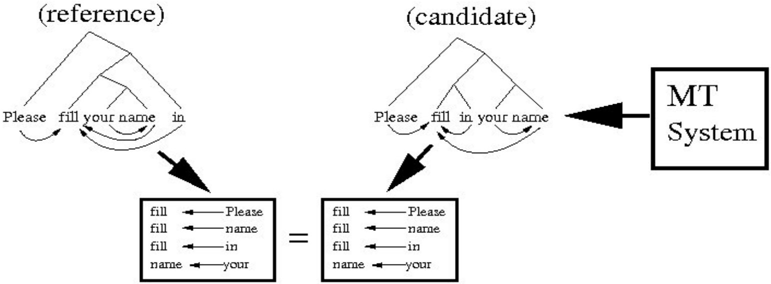
- Computes unigram precision and recall; penalises crossing alignments $\Rightarrow \gamma \cdot \left(\frac{\#\text{chunks}}{\#\text{unigram matches}}\right)^{\beta}$.
- But incorporates no notion of better or worse crossing alignments.

(Liu and Gildea, 2005) & (Owczarzak et al., 2007)



- Compare at the constituent or dependency level.
- Candidate is no longer punished for legitimate word-order variation.

(Liu and Gildea, 2005) & (Owczarzak et al., 2007)

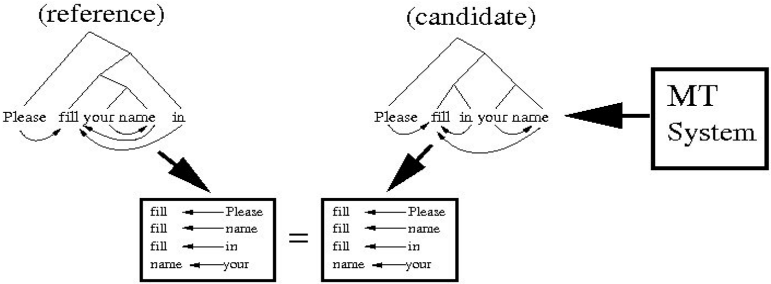


fill ← Please
fill ← name
fill ← in
name ← your

fill ← Please
fill ← name
fill ← in
name ← your

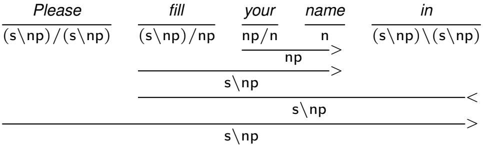
- Compare at the constituent or dependency level.
- Candidate is no longer punished for legitimate word-order variation.
- But: MT output is messy.

(Liu and Gildea, 2005) & (Owczarzak et al., 2007)

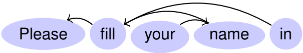


- Compare at the constituent or dependency level.
- Candidate is no longer punished for legitimate word-order variation.
- But: MT output is messy.
- How do you parse ill-formed input? (E.g., *Fill please your name in*.)

BLEUÂTRE: BLEU's Associate/Admirer(?) with Tectogrammatical RElations



⇓



⇓

∅	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{'fill'}
∅	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{'in','name'}
{'your'}	$\overleftarrow{left}$	**'name'**	$\overrightarrow{right}$	∅

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$

$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in' , 'name' }
{ 'your' }	$\overleftarrow{left}$	**' name '**	$\overrightarrow{right}$	$\emptyset$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$

$$\begin{array}{ccccc} \emptyset & \overleftarrow{left} & \textbf{`Please'} & \overrightarrow{right} & \{\text{ `fill' }\} \\ \emptyset & \overleftarrow{left} & \textbf{`fill'} & \overrightarrow{right} & \{\text{ `in' , `name' }\} \\ \{\text{ `your' }\} & \overleftarrow{left} & \textbf{` name '} & \overrightarrow{right} & \emptyset \end{array}$$

- **c2:** Please fill in your name

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$

$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in' , 'name' }
{ 'your' }	$\overleftarrow{left}$	' **name** '	$\overrightarrow{right}$	$\emptyset$

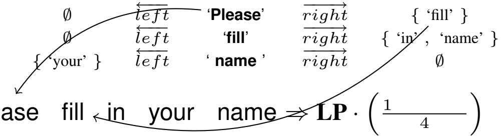
- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{\quad}{4} \right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in', 'name' }
{ 'your' }	$\overleftarrow{left}$	' **name** '	$\overrightarrow{right}$	$\emptyset$

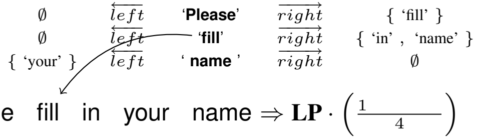
- **c2:** Please fill in your name ⇒ **LP** · $\left(\frac{1}{4}\right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in' , 'name' }
{ 'your' }	$\overleftarrow{left}$	**' name '**	$\overrightarrow{right}$	$\emptyset$

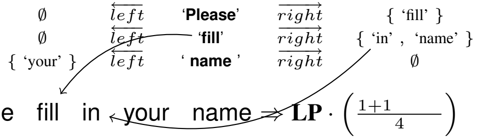
- **c2:** Please fill in your name ⇒ $\mathbf{LP} \cdot \left(\frac{1}{4} \right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp\left(1 - \frac{len(c)}{len(r)}\right), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$$\begin{array}{ccccc} \emptyset & \overleftarrow{left} & \textbf{`Please'} & \overrightarrow{right} & \{ \text{`fill'} \} \\ \emptyset & \overleftarrow{left} & \textbf{`fill'} & \overrightarrow{right} & \{ \text{`in'} , \text{`name'} \} \\ \{ \text{`your'} \} & \overleftarrow{left} & \textbf{` name '} & \overrightarrow{right} & \emptyset \end{array}$$

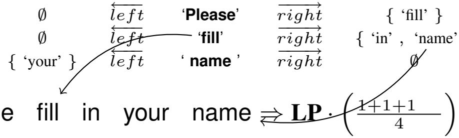
- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1}{4} \right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in' , 'name' }
{ 'your' }	$\overleftarrow{left}$	**' name '**	$\overrightarrow{right}$	$\emptyset$

- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{\cancel{1}+1+1}{4} \right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, & \text{if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), & \text{otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$

$$\begin{array}{ccccc} \emptyset & \overleftarrow{left} & \textbf{‘Please’} & \overrightarrow{right} & \{\ \text{‘fill’}\ \} \\ \emptyset & \overleftarrow{left} & \textbf{‘fill’} & \overrightarrow{right} & \{\ \text{‘in’}\ ,\ \text{‘name’}\ \} \\ \{\ \text{‘your’}\ \} & \overleftarrow{left} & \textbf{‘ name ’} & \overrightarrow{right} & \emptyset \end{array}$$

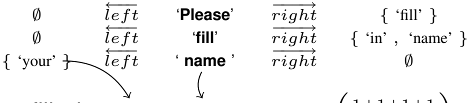
- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1}{4} \right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$$\begin{array}{ccccc} \emptyset & \overleftarrow{left} & \textbf{`Please'} & \overrightarrow{right} & \{\text{ `fill' }\} \\ \emptyset & \overleftarrow{left} & \textbf{`fill'} & \overrightarrow{right} & \{\text{ `in' , `name' }\} \\ \{\text{ `your' }\} & \overleftarrow{left} & \textbf{` name '} & \overrightarrow{right} & \emptyset \end{array}$$

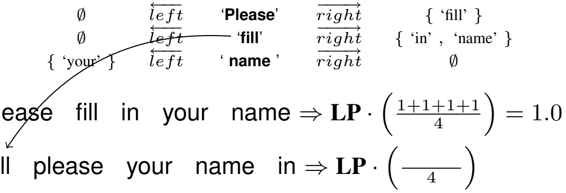
- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$$\begin{array}{ccccc} \emptyset & \overleftarrow{left} & \textbf{`Please'} & \overrightarrow{right} & \{\text{ `fill' }\} \\ \emptyset & \overleftarrow{left} & \textbf{`fill'} & \overrightarrow{right} & \{\text{ `in' , `name' }\} \\ \{\text{ `your' }\} & \overleftarrow{left} & \textbf{` name '} & \overrightarrow{right} & \emptyset \end{array}$$

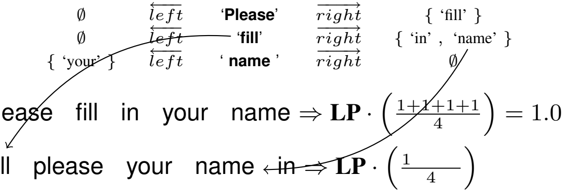
- **c2:** Please fill in your name ⇒ $\mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$
- **c1:** Fill please your name in ⇒ $\mathbf{LP} \cdot \left(\frac{}{4}\right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in' , 'name' }
{ 'your' }	$\overleftarrow{left}$	**' name '**	$\overrightarrow{right}$	$\emptyset$

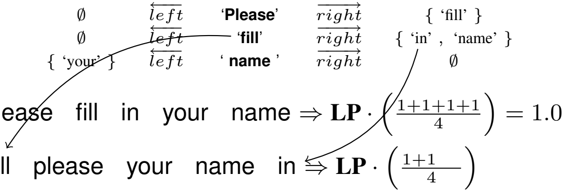
- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$
- **c1:** Fill please your name in $\Rightarrow \mathbf{LP} \cdot \left(\frac{1}{4}\right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in' , 'name' }
{ 'your' }	$\overleftarrow{left}$	' **name** '	$\overrightarrow{right}$	$\emptyset$

- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$
- **c1:** Fill please your name in $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1}{4}\right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$

$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{ 'fill' }
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{ 'in' , 'name' }
{ 'your' }	$\overleftarrow{left}$	**' name '**	$\overrightarrow{right}$	$\emptyset$

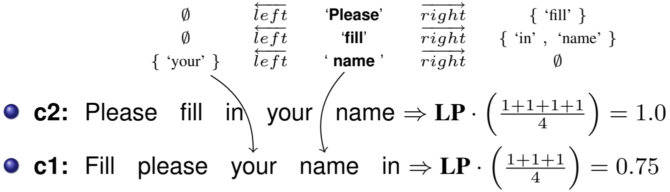
- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$
- **c1:** Fill please your name in $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1}{4}\right)$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$



$\emptyset$	$\overleftarrow{left}$	**‘Please’**	$\overrightarrow{right}$	{ ‘fill’ }
$\emptyset$	$\overleftarrow{left}$	**‘fill’**	$\overrightarrow{right}$	{ ‘in’ , ‘name’ }
{ ‘your’ }	$\overleftarrow{left}$	**‘ name ’**	$\overrightarrow{right}$	$\emptyset$

- **c2:** Please fill in your name ⇒ $\mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$
- **c1:** Fill please your name in ⇒ $\mathbf{LP} \cdot \left(\frac{1+1+1}{4}\right) = 0.75$

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$

$$\begin{array}{ccccc} \emptyset & \overleftarrow{left} & \textbf{`Please'} & \overrightarrow{right} & \{\ \text{`fill'}\ \} \\ \emptyset & \overleftarrow{left} & \textbf{`fill'} & \overrightarrow{right} & \{\ \text{`in'}\ ,\ \text{`name'}\ \} \\ \{\ \text{`your'}\ \} & \overleftarrow{left} & \textbf{` name '} & \overrightarrow{right} & \emptyset \end{array}$$

- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$
- **c1:** Fill please your name in $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1}{4}\right) = 0.75$
- Well-formed candidate no longer penalised, **and** ill-formed candidate **is** penalised.

BLEUÂTRE: How it works

$$\text{BLEUÂTRE}_{c,r} = LengthPen \cdot \text{RECALL-OF-PARTIAL-ORDERINGS}$$

where:

$$LengthPen_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ exp(1 - \frac{len(c)}{len(r)}), \text{ otherwise} \end{cases} = \text{OPPOSITE OF BLEU's BP}$$

$$\begin{array}{ccccc}
\emptyset & \overleftarrow{left} & \textbf{`Please'} & \overrightarrow{right} & \{\ \text{`fill'}\ \} \\
\emptyset & \overleftarrow{left} & \textbf{`fill'} & \overrightarrow{right} & \{\ \text{`in'}\ ,\ \text{`name'}\ \} \\
\{\ \text{`your'}\ \} & \overleftarrow{left} & \textbf{` name '} & \overrightarrow{right} & \emptyset
\end{array}$$

- **c2:** Please fill in your name $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1+1}{4}\right) = 1.0$
- **c1:** Fill please your name in $\Rightarrow \mathbf{LP} \cdot \left(\frac{1+1+1}{4}\right) = 0.75$
- Well-formed candidate no longer penalised, **and** ill-formed candidate **is** penalised.
- Even unparsable (or unreliably parsable) strings can be scored.

TIDES MTC (2 & 4): Comparison with (Owczarzak et al., 2007)

Fluency		Accuracy		Ave.	
Bleu	0.155*	**Meteor**	0.278*	**Meteor**	0.242*
Ow. et al.	0.154*	**NIST**	0.273*	**NIST**	0.238*
Meteor	0.149*	**GTM**	0.260*	**Ow. et al.**	0.236*
NIST	0.146*	**Ow. et al.**	0.224*	**GTM**	0.230*
GTM	0.146*	**BA**	0.202	**Bleu**	0.197*
TER	-0.133*	**Bleu**	0.199*	**BA**	0.186
BleuÂtre (BA)	0.128	**TER**	-0.192*	**TER**	-0.182*

Table: Correlation to human judgments. (**GTM**=Generalised Text Matcher; **TER**=Translation Edit Rate.) (Difference of ± 0.015 is significant at 95%. (* $=$ results are as reported in (Owczarzak et al., 2007).)

- (Owczarzak et al., 2007) use LFG dependency triples (here pred-arg only) — compute f-score of candidate.
- BleuÂtre on a par with TER and (sometimes) Bleu.

BLEUÂTRE vs. Direct Syntax-based Approach: We *Can* Simulate Parsing

FLUENCY		**ACCURACY**		**AVE.**	
Unlab. F-score (UFS)	0.143	BA	0.208	BA	0.190
Lab. F-score (LFS)	0.142	UFS	0.196	UFS	0.189
BLEUÂTRE (BA)	0.130	LFS	0.194	LFS	0.188

Table: Pearson's correlation between BLEUÂTRE, and C&C parser-based f-score evaluation (labelled and unlabelled). Only a difference of ± 0.016 is significant with 95% confidence.

- MTC Sections 2 and 4 (only 14,138 judgment-reference-score triples due to parsing errors).
- Differences are not significant ⇒ BLEUÂTRE and direct syntax-based approach (with same parser and grammatical dep's — C&C) are the same.

BLEUÂTRE vs. METEOR (v 0.5)

	BLEUÂTRE	METEOR
E09	0.338	0.351
E11	0.193	0.253
E12	0.216	0.264
E14	0.257	0.285
E15	0.238	0.237
E22	0.273	0.284
AVE	0.253	0.279

Table: BLEUÂTRE and METEOR's correlation (no stemming or WordNet) to an average of human judgments of fluency and accuracy for various MT systems. ± 0.016 is significant at 95% ($p \leq$ 3.609e-11.)

- BLEUÂTRE and METEOR use all 4 reference translations. (BLEUÂTRE score is best single comparison to a reference.)
- Performances do not always differ significantly (only slightly in the average).

BLEUÂTRE vs. (Liu and Gildea, 2005)

E14-FLUENCY		E15-FLUENCY	
BLEUÂTRE	0.199	**BLEUÂTRE**	0.188
LG_dt	0.159*	**LG_pt**	0.144*
LG_dc	0.157*	**LG_dt**	0.137*
LG_pt	0.147*	**LG_dc**	0.128*
BLEU	0.132*	**BLEU**	0.122*
LG_dtvc	0.090*	**LG_ptvc**	0.089*
LG_ptvc	0.065*	**LG_dtvc**	0.066*

Table: Correlation of BLEUÂTRE and Liu and Gildea's metrics to human fluency judgments. (Key: * indicates that the score is from (Liu and Gildea, 2005); **LG**=Liu and Gildea — different approaches: **_dt**=dependency subtrees, **vc**=vector-cosines, **_pt** structural subtrees; **_dc**=dependency chains.) ± 0.06 difference is significant with 95% confidence (by our calculations).

- Same data set (*modulo* 1% parsing failures).
- BLEUÂTRE perhaps outperforms more complex use of parses.
- Are performance differences due to methodological (BLEUÂTRE vs. their approaches), or parser- and grammar-based reasons?

BLEUÂTRE on MTC 2 and 4, Multiple References

FLUENCY	ACCURACY	AVE.
0.235	0.328	0.315

Table: BLEUÂTRE correlation to across-judge (average of individual) human judgments using multiple references (MTC 2 and 4). ± 0.015 significant at 95%.

- BLEUÂTRE meta-evaluation results for entire MTC (2 and 4) with multiple references.
- For comparison: no similar figures reported by other authors (to our knowledge).

Conclusions and Future Work

- Simulating parsing in MT eval. *is* possible ⇒ holding parser and grammar constant.
- Performance better than some syntax-based results, worse than others. ⇒ Suspect nature of dependencies as cause of low performance w.r.t. (Owczarzak et al., 2007).
- With access to multiple reference translations, BLEUÂTRE and METEOR (v 0.5, no stemming or WordNet) are comparable.
- Future work:
  - Incorporate "soft matching" (WordNet), and automatic paraphrase-generating techniques.
  - Add NIST-like "informativeness" weights to flattened dep's
  - Perform more direct, full-featured comparison between BLEUÂTRE and Ow. et al., METEOR, etc.
- Thank you for your attention.

Banerjee, S. and Lavie, A. 2005.
METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.
In *Proceedings the ACL*, Ann Arbor, MI, USA.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. 2007.
Open information extraction from the web.
In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Brew, C. and Thompson, H. S. 1994.
Automatic evaluation of computer generated text: a progress report on the TextEval project.
In *Proceedings of the Workshop on Human Language Technology*, pages 108–113.

Callison-Burch, C. M., Osborne, M., and Koehn, P. 2006.
Re-evaluating the role of BLEU in machine translation research.

In *Proceedings of the EACL-2006*, Trento, Italy.

Liu, D. and Gildea, D. 2005.
Syntactic features for evaluation of machine translation.
In *the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA.

Owczarzak, K., van Genabith, J., and Way, A. 2007.
Dependency-based automatic evaluation for machine translation.
In *Proceedings of the HLT-NAACL Workshop on Syntax and Structure in Statistical Translation*, Rochester, NY, USA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002.
BLEU: a method for automatic evaluation of machine translation.
In *Proceedings of the ACL*, Philadelphia, PA, USA.

Thompson, H. 1991.
Automatic evaluation of translation quality: Outline of methodology

and report on pilot experiment.
In *(ISSCO) Proceedings of the Evaluators Forum*, Geneva, Switzerland.