A Walk on the Other Side:
Adding Statistical Components to a Transfer-Based Translation System

Ariadna Font Llitjós & Stephan Vogel
{aria,vogel+}@cs.cmu.edu

Goal

Complement the current trend of adding more structure to Statistical Machine Translation systems by exploring the opposite direction: adding statistical components to a Transfer-Based MT system.
- Statistical phrase alignment for semiautomatic lexicon generation
- Minimum error rate training
- *N*-best list rescoring with statistical features

Hybrid Transfer-Based (Xfer) system incorporates:

- *N*-gram LM
- Fragmentation weight (least rules to cover most of the output are preferred)
- Length penalty

Data: Basic Travel Expression Corpus (BTEC)

		Train	Test
BTEC	Sentence Pairs	123,416	506
	Word Tokens	903,525	3,764
	Word Types	12,578	776
	Coverage	--	756 (97%)

Grammar and Lexicon (Baseline vs. Refined)

-Refined Grammar and Lexicon are the result of automatic rule expansion and Improvement (Font Llitjos and Ridmann 2007): a total of 8 translation rules and 30 constraints were added.

```
{NP,8}
NP::NP : [DET ADJ N] → [DET N ADJ]
;; alignments
(  (X1::Y1) (X2::Y3) (X3::Y2)
;; analysis constraints      ;; transfer constraints      ;; generation constraints
   ((x0 det) = x1)              (y0 = x0)                    ((y1 agr num) = (y2 agr num))
   ((x0 mod) = x2)              (y1 == (y0 det))              ((y1 agr gen) = (y2 agr gen))  )
   (x0 = x3)                    (y3 == (y0 mod))
                                (y2 = y0)
                                ((y1 agr) = (x1 agr))
```

- BTEC Lexicon was semi-automatically augmented to adapt for the new domain (from 474 to 1,732 lexical entries):
    - Trained IBM1 and extracted phrases with tight pruning
    - Manually annotated POS and feature constraints

Initial Results: Lower and Upper Bound

	Systems	METEOR	BLEU	NIST
No Ranking (1st output)	Baseline	0.5666	0.2745	5.88
	Refined	0.5676	0.2559	5.62
Initial Ranking (1st best)	Baseline	0.6176	0.3425	6.53
	Refined	0.6222	0.3513	6.56
Ideal Ranking (Auto Oracle)	Baseline	0.6863	0.4068	7.42
	Refined	0.6954	0.4215	7.51

Adding Statistical Components to the Ranker:

- Word-to-word probabilities

$$P(e \mid s) = \frac{1}{J^I} \prod \sum p(e_i \mid s_j) \qquad P(s \mid e) = \frac{1}{I^J} \prod \sum p(s_j \mid e_i)$$

- Conditional Rule probabilities given Rule Type

$$P(D) = \prod p(r \mid R)$$

- *N*-gram LM for Rules and Rule Types (POS LM)

$$P(D) = \prod p(r \mid r_{-n} \ldots r_{-1}) \qquad P(D) = \prod p(R \mid R_{-n} \ldots R_{-1})$$

MER training (BLEU) for Refined System output (Venugopal)

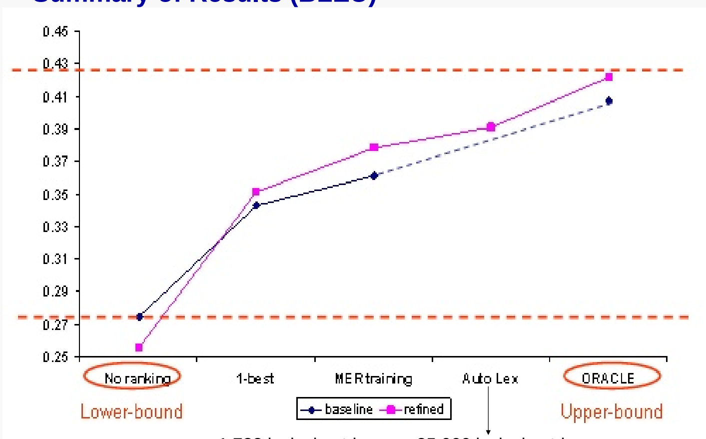
	System + Statistical Components	1-best BLEU
Rule Based	Xfer	0.2559
+ Stat. Comp.	Xfer + LM + Frag	0.3513
Optimal weights	LM + Rule Type LM	0.3736
	LM + Rule Type + Rule LM	0.3744
	LM + Frag + Len + Rule Type LM + Rule Prob.	0.3746

Xfer with Optimal weights (LM, Frag)

System	METEOR	BLEU	NIST
Baseline	0.6184	0.3609	6.68
Refined	0.6231	0.3780	6.79

p = 0.0051

Summary of Results (BLEU)



1,732 lexical entries   25,000 lexical entries

Example Translations
with (WO) and without (NO) weight optimization

1 Src: where is the boarding gate ? NO: dónde está *el embarque puerta* ? WO: dónde está <u>la puerta embarque</u> ?
2 Src: i would like a twin room with a bath please . NO: me gustaría habitación *una* cama doble con un baño por favor . WO: me gustaría <u>una</u> habitación cama doble con un baño por favor .
3 Src: does he speak japanese ? NO: él *hablar a* japonés ? WO: <u>habla</u> japonés ?
4 Src: do you sell duty-free items ? NO: *te* venden artículos duty-free ? WO: <u>vendéis</u> artículos duty-free ?

Conclusions

- **Word and phrase alignment techniques** allows to **quickly augment** the Xfer **lexicon**.
- When selecting good translations from *n*-best lists, **most gain comes** from the **Statistical LM**, which was already part of the Xfer system.
- Adding **additional features**, such as word-to-word probabilities and rule (type) probabilities, further **improves performance**.
- **MER training** becomes **crucial** when **multiple components** are used in the decoder

Carnegie Mellon