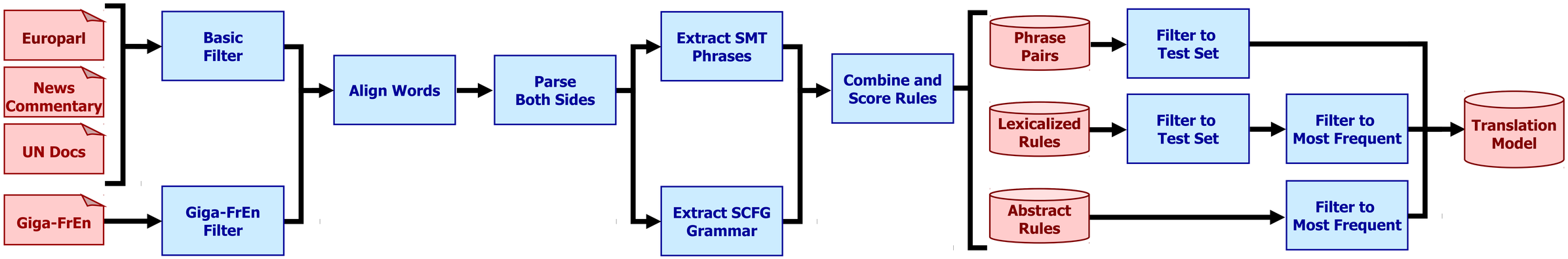
CMU Syntax-Based Machine Translation at WMT 2011

Carnegie Mellon University

Greg Hanneman and Alon Lavie
{ghannema, alavie}@cs.cmu.edu

Language Technologies Institute, Carnegie Mellon University



Training Data Selection

- Basic Filtering: Keep balanced lines of ≤ 95 words with all words ≤ 25 characters
- Giga-FrEn Filtering: Keep balanced lines of ≤ 50 words containing a French word seen < 20 times
- Number of parallel sentences:

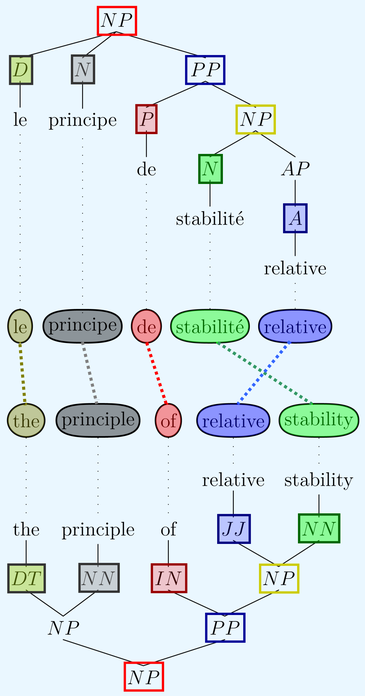
Corpus	Released	Used
Europarl	1,825,077	1,614,111
News Comm	115,562	95,138
UN Docs	12,317,600	9,352,232
Giga-FrEn	22,520,400	2,839,466
Total	**38,778,639**	**13,900,947**

- **+1.2 to +1.9 BLEU** and **–45% OOV rate** over using WMT 2010 data only

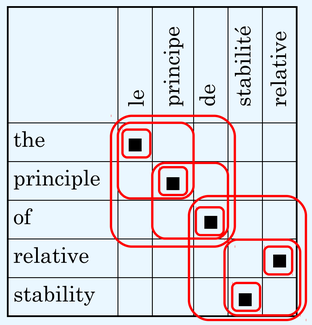
External Tools

- MGIZA++
- Stanford tokenizer
- Berkeley parser
- Joshua decoder
- SRI LM toolkit
- ZMERT

Grammar Extraction



- Extract SCFG rules as well as standard SMT phrase pairs
- When duplicates occur, keep syntactic version but share counts
- Give SMT phrase pairs dummy syntactic label



N::NN → [stabilité]::[stability]
NP::NP → [stabilité relative]::[relative stability]
NP::NP → [N¹ A²]::[JJ² NN¹]
PHR::PHR → [principe de]::[principle of]

Grammar Filtering

- Always keep all phrase pairs matching test set
- Method 1: Keep the 10,000 most frequently extracted hierarchical rules of all types
- Method 2: Keep only the 2000 most frequent abstract rules, but the 100,000 most frequent partially lexicalized rules

Results

- Score differences between grammar filtering techniques inconclusive (**–0.4 to +0.8 BLEU**)
- Method 1 uses more rule instances; Method 2 uses more unique rules
- Automatic metric scores on dev test sets:

	newstest-2009		newstest-2010	
System	**METR**	**BLEU**	**METR**	**BLEU**
WMT '10 Method 1	54.94	24.77	56.66	25.78
WMT '10 Method 2	55.16	24.88	56.89	26.05
WMT '11 Method 1	55.82	26.02	58.13	27.71
WMT '11 Method 2	55.77	26.01	57.88	27.38