ITG for Joint Phrasal Translation Modeling

Colin Cherry
University of Alberta

Dekang Lin
Google Inc.

The Gist

- Joint phrasal translation models (JPTM) learn a bilingual phrase table using EM
- Phrasal ITG:
  - Use synchronous parsing to replace **hill climbing** & **sampling** with dynamic programming
- Do resulting phrase tables improve translation?

Outline

- Phrasal Translation Models
- We build on:
  - Phrase extraction, JPTM, ITG
- Phrasal ITG
  - Helpful constraints
- Results
- Summary & Future Work

Phrasal translation model

English	French	P(e|f)	P(f|e)
ethical food	alimentation éthique	0.95	0.16
ethical foreign policy	politique étrangère morale	0.23	0.01
ethical foundations	fondements éthiques	0.10	0.03
…			

- Ultimately interested in a bilingual phrase table
  - Lists and scores possible phrasal translations

Surface Heuristic

cars				●	
red					●
likes		●			
he	●				
	il	aime	les	voitures	rouges

- Alignments provided by GIZA++ combination
- Surface heuristic:
  - Count each consistent phrase as occurring once
  - Aggregate counts over all sentence pairs

Surface Heuristic

cars				●	
red					●
likes		●			
he	●				
	il	aime	les	voitures	rouges

- Alignments provided by GIZA++ combination
- Surface heuristic:
  - Count each consistent phrase as occurring once
  - Aggregate counts over all sentence pairs

Surface Heuristic

cars				●	
red					●
likes		●			
he	●				
	il	aime	les	voitures	rouges

- Alignments provided by GIZA++ combination
- Surface heuristic:
  - Count each consistent phrase as occurring once
  - Aggregate counts over all sentence pairs

Surface Heuristic

cars				●	
red					●
likes		●			
he	●				
	il	aime	les	voitures	rouges

- Alignments provided by GIZA++ combination
- Surface heuristic:
  - Count each consistent phrase as occurring once
  - Aggregate counts over all sentence pairs

Joint Phrasal Model (JPTM)

- Introduced by Marcu and Wong (2002)
- Trained with EM, like the IBM models
- Sentence pair built simultaneously
  - Generate a bag of bilingual phrase pairs
  - Permute the phrases to form *e* and *f*

$$P(e,f) \propto \sum_{A} \left[\prod_{(\bar{e_i},\bar{f_i}) \in A} p(\bar{e_i},\bar{f_i}) \right]$$

Joint Phrasal Model

cars					
red					
likes					
he					
	il	aime	les	voitures	rouges

Reason over an exponential number of phrasal alignments

Space is huge - task actually accomplished by sampling around high-probability point

Joint Phrasal Model

cars					
red					
likes					
he					
	il	aime	les	voitures	rouges

Reason over an exponential number of phrasal alignments

Space is huge - task actually accomplished by sampling around high-probability point

Joint Phrasal Model

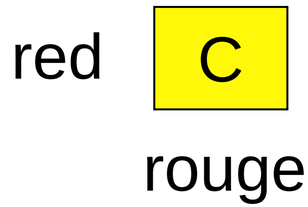
cars				●	
red					●
likes		●			
he	●				
	il	aime	les	voitures	rouges

Birch et al. (2006): Constrained JPTM

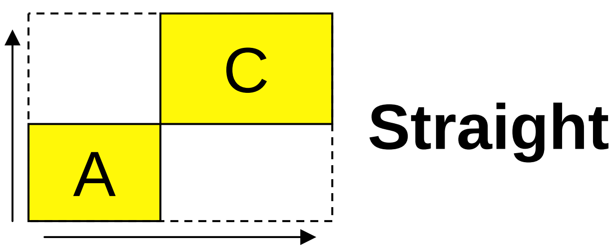
Explore only phrasal alignments consistent with high precision word alignment

Inversion Transduction Grammar

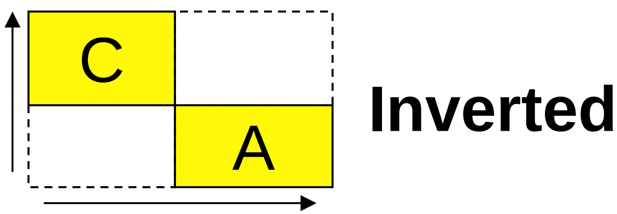
- Introduced in by Wu (1997)
  - Transduction:
    - C $\rightarrow$ red / rouge



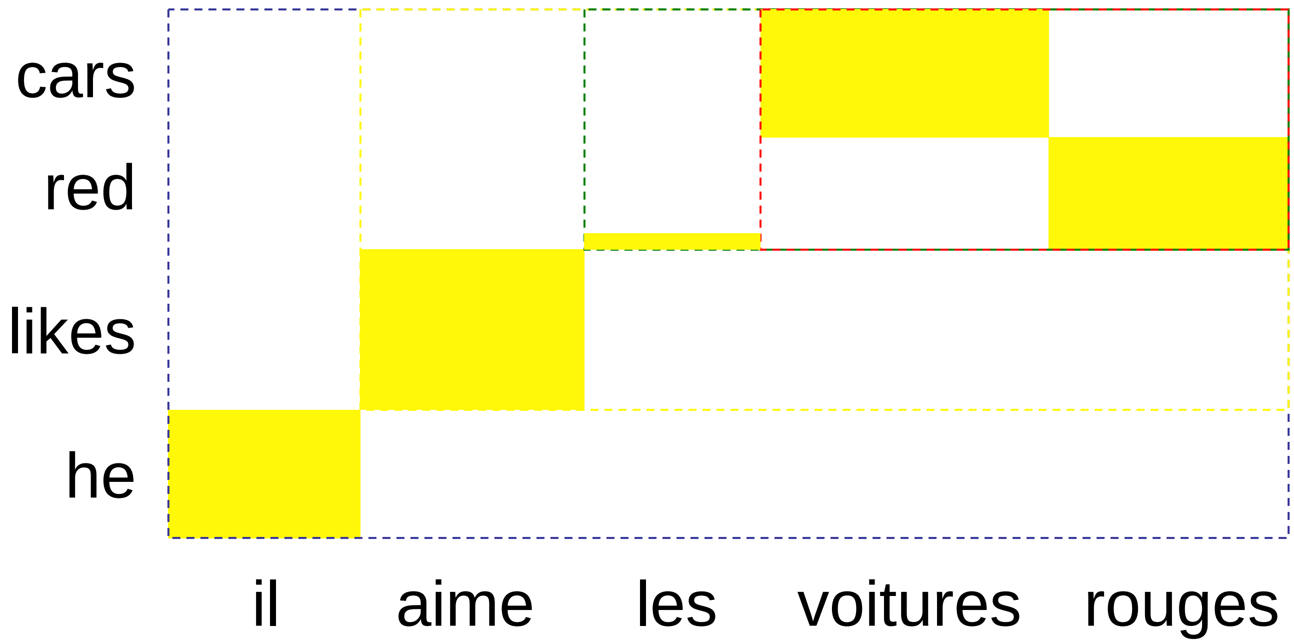
  - Inversion:
    - A $\rightarrow$ [A C]



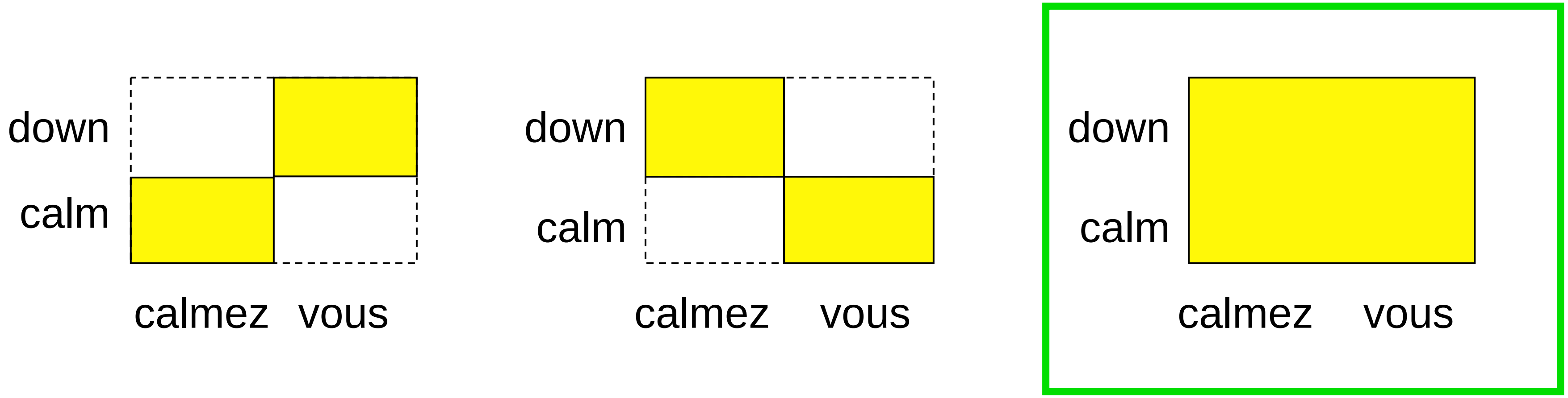
    - B $\rightarrow$ <A C>



ITG Parse



Phrasal ITG



- Any phrase pair can be produced by the lexicon
- Choose between straight, inverted and now: **phrasal**

Training Phrasal ITG

$$C \rightarrow \bar{e}/\bar{f} \text{ with probability } P\left(\bar{e}/\bar{f}|C\right)$$

- All phrase pairs share mass as a joint model
- Can be trained unsupervised with inside-outside
- No more expensive than binary bracketing:
  - Phrases were already being explored as constituents

The hope

- By moving to exact expectation:
  - Create more accurate statistics
  - Find a larger variety of phrase pairs

The problem - still slow: $O(n^6)$

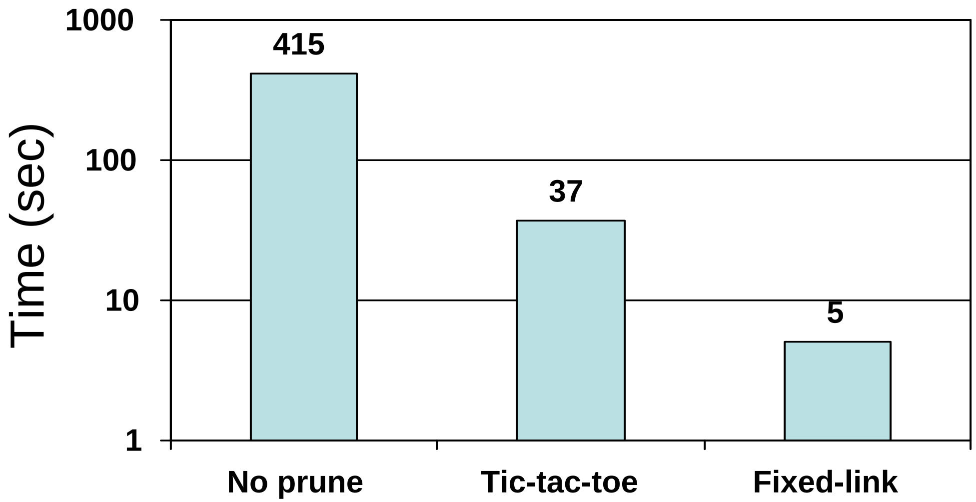
- ITG algorithms can be pruned:
  - $O(n^4)$ potential constituents are considered
  - $O(n^2)$ time spent considering all ways to build each constituent

- **Fixed link pruning**: Eliminate constituents that are not consistent with a given word alignment
  - Skip them and treat them as having 0 probability

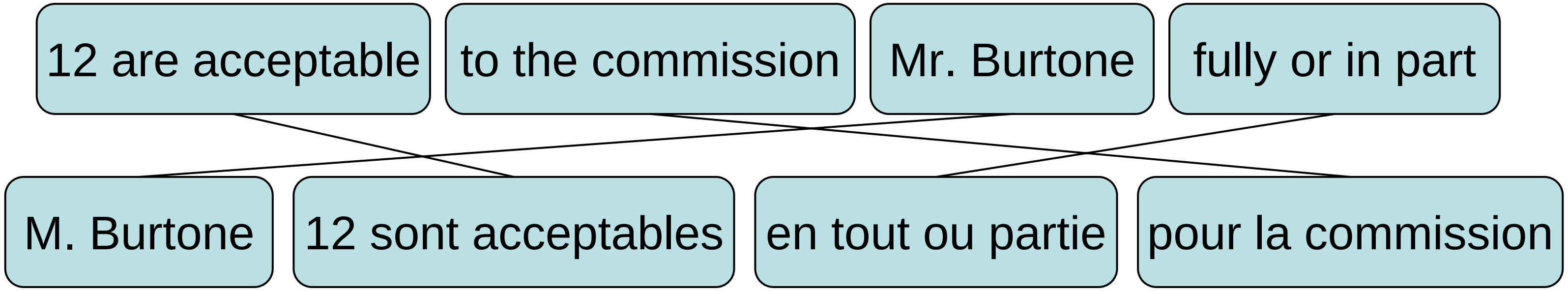
- One link can potentially rule out 50% of constituents

Fixed Link Speed-up

- Used GIZA++ intersection alignments
- Inside-outside on first 100 sentences of corpus
- Compared to Tic-tac-toe (Zhang & Gildea 2005)



What about the ITG constraint?

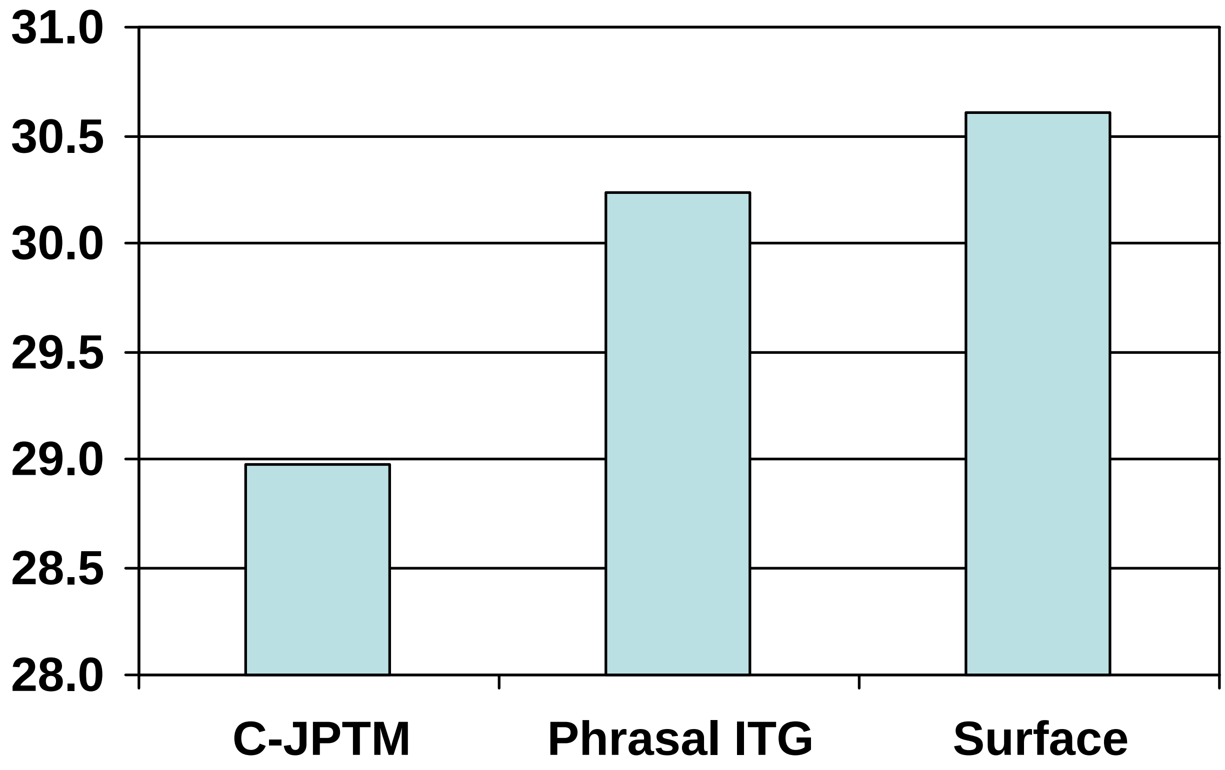


- ITG can’t handle this due to discontinuous constituents
- Check fixed links used for pruning
  - If they are non-ITG, drop from training set
- In our French-English Europarl set, this results in a reduction in data of less than 1%

Experiments

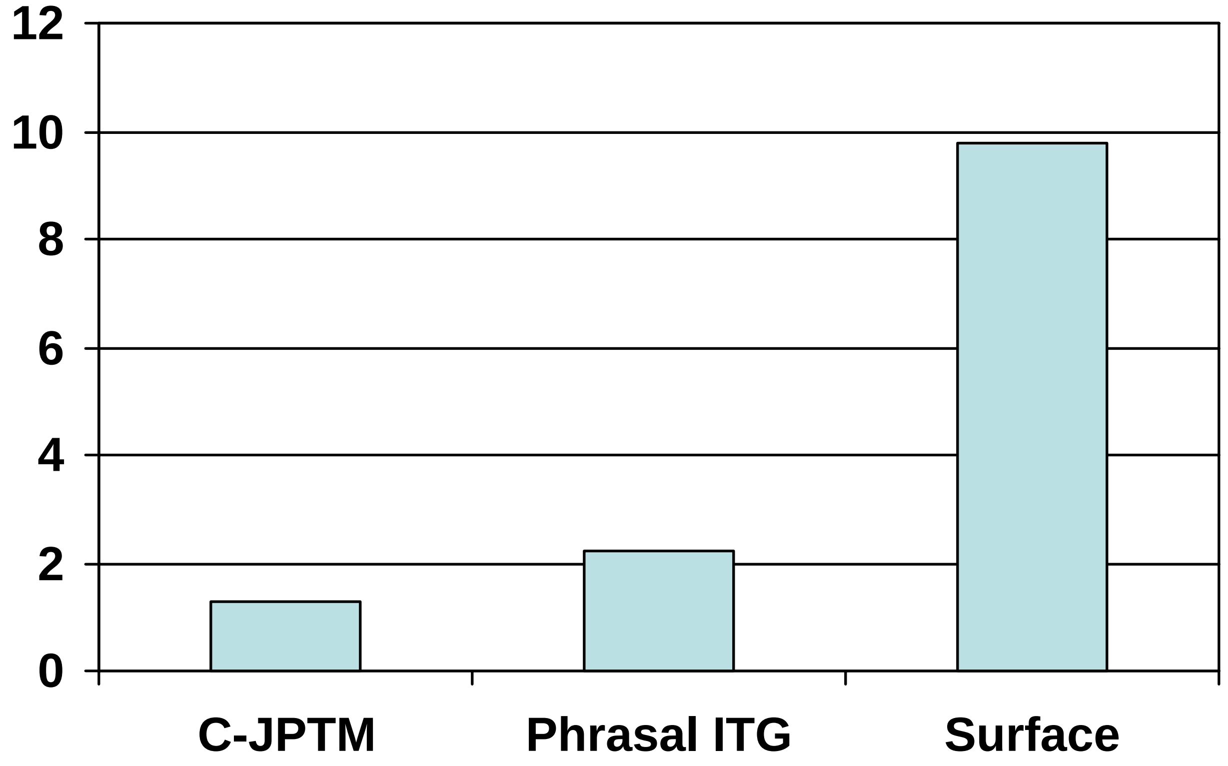
- Conditionalize joint tables to P(e|f) and P(f|e)
- French-English Europarl Set
  - 25 length limit, 400k sentence pairs
- SMT Workshop Baseline MT System
  - Pharaoh, MERT Training on 500 tuning pairs
- Included unnormalized IBM Model 1 features for all
- Compared to:
  - JPTM constrained with GIZA++ Intersect
  - Surface Heuristic Extraction with GIZA++ GDF

Results: BLEU Scores



Results: Table Size

(in millions of entries)



Summary

- Phrasal ITG that learns phrases from bitext
  - Similar to JPTM
- Complete expectations do matter
  - Other JPTMs could benefit from improving their search and sampling methods
- A new ITG pruning technique
  - 80 times faster inside-outside

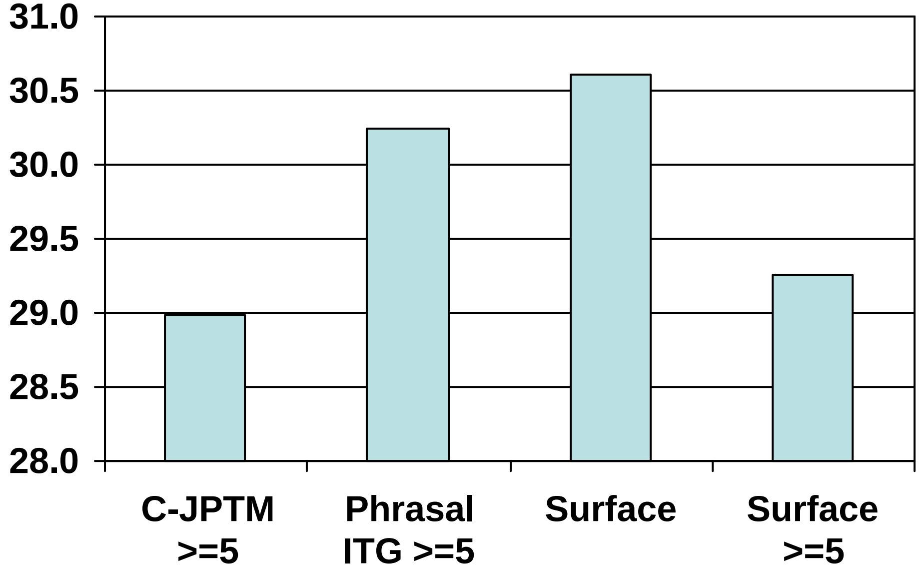
Future: Eliminate Frequency Limits

- Must constrain any joint model to use phrases that occur with a minimum frequency
  - Otherwise sentence = phrase is ML solution

cars					
red					
likes					
he					
	il	aime	les	voitures	rouges

Future: Eliminate Frequency Limits

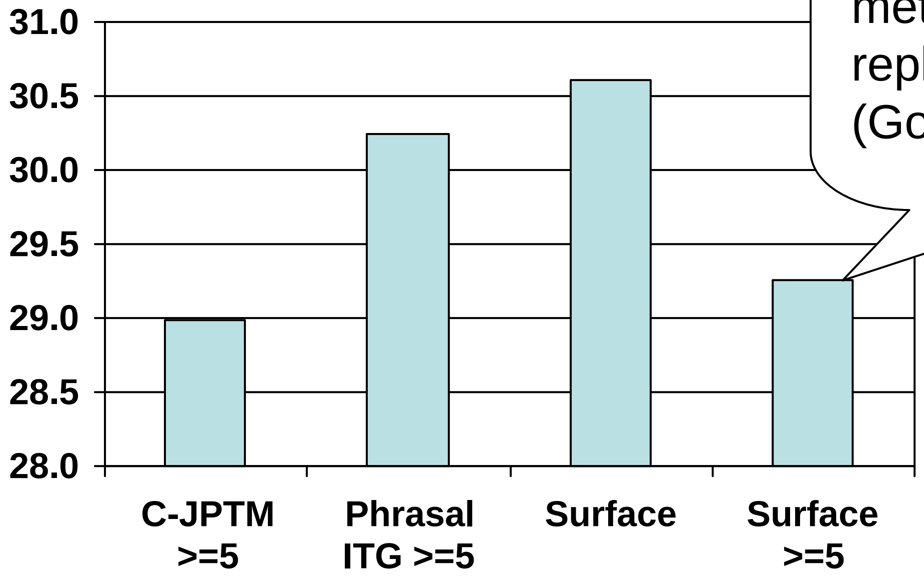
- Must constrain any joint model to use phrases that occur with a minimum frequency
  - Otherwise sentence = phrase is ML solution



Future: Eliminate Frequency Limits

- Must constrain any joint model to use phrases that occur with a minimum frequency
  - Otherwise sentence = phrase is ML

Apply Bayesian methods (priors) to replace these limits (Goldwater et al. 2006)



This isn't the whole story…

- Explored the same model as a **phrasal aligner**
- Needs additional constraints to work:
  - Fixed links help select phrases that are non-compositional
- Alignments work well with surface heuristic
- Details in the paper!

Questions? Comments? Suggestions?

Support provided by:

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Alberta Ingenuity Fund

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Alberta Informatics Circle of Research Excellence

Along the way…

- Adapt consistency constraints from heuristic phrase extraction for ITG parsing
- Deal with the ITG constraint in large data