
Alon Lavie
Language Technologies Institute
Carnegie Mellon University

Joint work with:
Greg Hanneman, Vamshi Ambati, Alok Parlikar, Edmund Huber,
Jonathan Clark, Erik Peterson, Christian Monson, Abhaya Agarwal,
Kathrin Probst, Ari Font Llitjos, Lori Levin, Jaime Carbonell, Bob
Frederking, Stephan Vogel
Outline

• Context and Rationale
• CMU Statistical Transfer MT Framework
• Extracting Syntax-based MT Resources from Parallel-corpora
• Integrating Syntax-based and Phrase-based Resources
• Open Research Problems
• Conclusions
Rule-based vs. Statistical MT

• Traditional Rule-based MT:
  – Expressive and linguistically-rich formalisms capable of describing complex mappings between the two languages
  – Accurate “clean” resources
  – Everything constructed manually by experts
  – Main challenge: obtaining and maintaining broad coverage

• Phrase-based Statistical MT:
  – Learn word and phrase correspondences automatically from large volumes of parallel data
  – Search-based “decoding” framework:
    • Models propose many alternative translations
    • Effective search algorithms find the “best” translation
  – Main challenge: obtaining and maintaining high translation accuracy
Research Goals

• Long-term research agenda (since 2000) focused on developing a unified framework for MT that addresses the core fundamental weaknesses of previous approaches:
  - Representation – explore richer formalisms that can capture complex divergences between languages
  - Ability to handle morphologically complex languages
  - Methods for automatically acquiring MT resources from available data and combining them with manual resources
  - Ability to address both rich and poor resource scenarios

• Main research funding sources: NSF (AVENUE and LETRAS projects) and DARPA (GALE)
CMU Statistical Transfer (Stat-XFER) MT Approach

- Integrate the major strengths of rule-based and statistical MT within a common framework:
  - Linguistically rich formalism that can express complex and abstract compositional transfer rules
  - Rules can be written by human experts and also acquired automatically from data
  - Easy integration of morphological analyzers and generators
  - Word and syntactic-phrase correspondences can be automatically acquired from parallel text
  - Search-based decoding from statistical MT adapted to find the best translation within the search space: multi-feature scoring, beam-search, parameter optimization, etc.
  - Framework suitable for both resource-rich and resource-poor language scenarios
Stat-XFER Main Principles

• Framework: Statistical search-based approach with syntactic translation transfer rules that can be acquired from data but also developed and extended by experts
• Automatic Word and Phrase translation lexicon acquisition from parallel data
• Transfer-rule Learning: apply ML-based methods to automatically acquire syntactic transfer rules for translation between the two languages
• Elicitation: use bilingual native informants to produce a small high-quality word-aligned bilingual corpus of translated phrases and sentences
• Rule Refinement: refine the acquired rules via a process of interaction with bilingual informants
• XFER + Decoder:
  – XFER engine produces a lattice of possible transferred structures at all levels
  – Decoder searches and selects the best scoring combination
Stat-XFER MT Approach

Interlingua

Semantic Analysis

Sentence Planning

Syntactic Parsing

Transfer Rules

Statistical-XFER

Source (e.g. Arabic)

Direct: SMT, EBMT

Target (e.g. English)

Text Generation
Stat-XFER Framework

Source Input

Preprocessing

Morphology

Transfer Engine

Transfer Rules

Bilingual Lexicon

Translation Lattice

Second-Stage Decoder

Language Model

Weighted Features

Target Output

1/21/2009

Alon Lavie: Stat-XFER
Transfer Rules

\{NP1,3\}
NP1::NP1 [NP1 "H" ADJ] -> [ADJ NP1]
((X3::Y1)
(X1::Y2)
((X1 def) = +)
((X1 status) = c absolute)
((X1 num) = (X3 num))
((X1 gen) = (X3 gen))
(X0 = X1))

Translation Lexicon

N::N |: ["$WR"] -> ["BULL"]
((X1::Y1)
((X0 NUM) = s)
((Y0 lex) = "BULL"))

N::N |: ["$WRH"] -> ["LINE"]
((X1::Y1)
((X0 NUM) = s)
((Y0 lex) = "LINE"))

Preprocessing

Source Input
בשורה הבאה

Morphology

Transfer Engine

Translation Output Lattice

(0 1 "IN" @PREP)
(1 1 "THE" @DET)
(2 2 "LINE" @N)
(1 2 "THE LINE" @NP)
(0 2 "IN LINE" @PP)
(0 4 "IN THE NEXT LINE" @PP)

Decoder

Language Model + Additional Features

English Output
in the next line
Transfer Rule Formalism

Type information
Part-of-speech/constituent information
Alignments
x-side constraints
y-side constraints
xy-constraints,
  e.g. ((Y1 AGR) = (X1 AGR))

;SL: the old man, TL: ha-ish ha-zaqen

NP::NP  [DET ADJ N] -> [DET N DET ADJ]

((X1::Y1)
(X1::Y3)
(X2::Y4)
(X3::Y2)

((X1 AGR) = *3-SING)
((X1 DEF = *DEF)
((X3 AGR) = *3-SING)
((X3 COUNT) = +)

((Y1 DEF) = *DEF)
((Y3 DEF) = *DEF)
((Y2 AGR) = *3-SING)
((Y2 GENDER) = (Y4 GENDER))
)
Transfer Rule Formalism

;SL: the old man, TL: ha-ish ha-zaqen

NP::NP   [DET ADJ N] -> [DET N DET ADJ]
  (X1::Y1)
  (X1::Y3)
  (X2::Y4)
  (X3::Y2)

  ((X1 AGR) = *3-SING)
  ((X1 DEF = *DEF)
  ((X3 AGR) = *3-SING)
  ((X3 COUNT) = +)

  ((Y1 DEF) = *DEF)
  ((Y3 DEF) = *DEF)
  ((Y2 AGR) = *3-SING)
  ((Y2 GENDER) = (Y4 GENDER))

Value constraints

Agreement constraints
Translation Lexicon: Hebrew-to-English Examples (Semi-manually-developed)

PRO::PRO |: ["ANI"] -> ["I"]
(X1::Y1)
((X0 per) = 1)
((X0 num) = s)
((X0 case) = nom)

PRO::PRO |: ["ATH"] -> ["you"]
(X1::Y1)
((X0 per) = 2)
((X0 num) = s)
((X0 gen) = m)
((X0 case) = nom)

N::N |: ["$&H"] -> ["HOUR"]
(X1::Y1)
((X0 NUM) = s)
((Y0 NUM) = s)
((Y0 lex) = "HOUR")

N::N |: ["$&H"] -> ["hours"]
(X1::Y1)
((Y0 NUM) = p)
((X0 NUM) = p)
((Y0 lex) = "HOUR")
Translation Lexicon:
French-to-English Examples
(Automatically-acquired)

DET::DET |: ["le"] -> ["the"]
( (X1::Y1) )

Prep::Prep |: ["dans"] -> ["in"]
( (X1::Y1) )

N::N |: ["principes"] -> ["principles"]
( (X1::Y1) )

N::N |: ["respect"] -> ["accordance"]
( (X1::Y1) )

NP::NP |: ["le respect"] -> ["accordance"]
( )

PP::PP |: ["dans le respect"] -> ["in accordance"]
( )

PP::PP |: ["des principes"] -> ["with the principles"]
( )
Hebrew-English Transfer Grammar
Example Rules
(Manually-developed)

{NP1,2}
;;SL: $MLH ADWMH
;;TL: A RED DRESS

NP1::NP1 [NP1 ADJ] -> [ADJ NP1]
  (X2::Y1)
  (X1::Y2)
  ((X1 def) = -)
  ((X1 status) = c absolute)
  ((X1 num) = (X2 num))
  ((X1 gen) = (X2 gen))
  (X0 = X1)

{NP1,3}
;;SL: H $MLWT H ADWMWT
;;TL: THE RED DRESSES

NP1::NP1 [NP1 "H" ADJ] -> [ADJ NP1]
  (X3::Y1)
  (X1::Y2)
  ((X1 def) = +)
  ((X1 status) = c absolute)
  ((X1 num) = (X3 num))
  ((X1 gen) = (X3 gen))
  (X0 = X1)
French-English Transfer Grammar
Example Rules
(Automatically-acquired)

{PP,24691}
;;SL: des principes
;;TL: with the principles

PP::PP [“des” N] -> [“with the” N]
( (X1::Y1) )

{PP,312}
;;SL: dans le respect des principes
;;TL: in accordance with the principles

PP::PP [Prep NP] -> [Prep NP]
( (X1::Y1) (X2::Y2) )
The Transfer Engine

• Input: source-language input sentence, or source-language confusion network
• Output: lattice representing collection of translation fragments at all levels supported by transfer rules
• Basic Algorithm: “bottom-up” integrated “parsing-transfer-generation” chart-parser guided by the synchronous transfer rules
  – Start with translations of individual words and phrases from translation lexicon
  – Create translations of larger constituents by applying applicable transfer rules to previously created lattice entries
  – Beam-search controls the exponential combinatorics of the search-space, using multiple scoring features
The Transfer Engine

• Some Unique Features:
  – Works with either learned or manually-developed transfer grammars
  – Handles rules with or without unification constraints
  – Supports interfacing with servers for morphological analysis and generation
  – Can handle ambiguous source-word analyses and/or SL segmentations represented in the form of lattice structures
Hebrew Example
(From [Lavie et al., 2004])

• Input word: B$WRH

0 1 2 3 4
|--------B$WRH--------|-----B-----|$WR|--H--|
|-----B-----|$WR|--H--|
|--B--|--H--|--$WRH--|
|--B--|--H--|--$WRH--|
Hebrew Example
(From [Lavie et al., 2004])

Y0: ((SPANSTART 0) (SPANEND 4) (LEX B$WRH) (POS N) (GEN F) (NUM S) (STATUS ABSOLUTE))
Y1: ((SPANSTART 0) (SPANEND 2) (LEX B) (POS PREP))
Y2: ((SPANSTART 1) (SPANEND 3) (LEX $WR) (POS N) (GEN M) (NUM S) (STATUS ABSOLUTE))
Y3: ((SPANSTART 3) (SPANEND 4) (LEX $LH) (POS POSS))
Y4: ((SPANSTART 0) (SPANEND 1) (LEX B) (POS PREP))
Y5: ((SPANSTART 1) (SPANEND 2) (LEX H) (POS DET))
Y6: ((SPANSTART 2) (SPANEND 4) (LEX $WRH) (POS N) (GEN F) (NUM S) (STATUS ABSOLUTE))
Y7: ((SPANSTART 0) (SPANEND 4) (LEX B$WRH) (POS LEX))
XFER Output Lattice

(28 28 "AND" -5.6988 "W" "(CONJ,0 'AND')")
(29 29 "SINCE" -8.20817 "MAZ" "(ADVP,0 (ADV,5 'SINCE'))")
(29 29 "SINCE THEN" -12.0165 "MAZ" "(ADVP,0 (ADV,6 'SINCE THEN'))")
(29 29 "EVER SINCE" -12.5564 "MAZ" "(ADVP,0 (ADV,4 'EVER SINCE'))")
(30 30 "WORKED" -10.9913 "&BD" "(VERB,0 (V,11 'WORKED'))")
(30 30 "FUNCTIONED" -16.0023 "&BD" "(VERB,0 (V,10 'FUNCTIONED'))")
(30 30 "WORSHIPPED" -17.3393 "&BD" "(VERB,0 (V,12 'WORSHIPPED'))")
(30 30 "SERVED" -11.5161 "&BD" "(VERB,0 (V,14 'SERVED'))")
(30 30 "SLAVE" -13.9523 "&BD" "(NP0,0 (N,34 'SLAVE'))")
(30 30 "BONDSMAN" -18.0325 "&BD" "(NP0,0 (N,36 'BONDSMAN'))")
(30 30 "A SLAVE" -16.8671 "&BD" "(NP,1 (LITERAL 'A') (NP2,0 (NP1,0 (NP0,0 (N,34 'SLAVE')))))")
(30 30 "A BONDSMAN" -21.0649 "&BD" "(NP,1 (LITERAL 'A') (NP2,0 (NP1,0 (NP0,0 (N,36 'BONDSMAN')))))")
The Lattice Decoder

- Stack Decoder, similar to standard Statistical MT decoders
- Searches for best-scoring path of non-overlapping lattice arcs
- No reordering during decoding
- Scoring based on log-linear combination of scoring features, with weights trained using Minimum Error Rate Training (MERT)
- Scoring components:
  - Statistical Language Model
  - Bi-directional MLE phrase and rule scores
  - Lexical Probabilities
  - Fragmentation: how many arcs to cover the entire translation?
  - Length Penalty: how far from expected target length?
ON THE FOURTH DAY THE LION ATE THE RABBIT TO A MORNING MEAL
Overall: -8.18323, Prob: -94.382, Rules: 0, Frag: 0.153846, Length: 0,
Words: 13,13

235 < 0 8 -19.7602: B H IWM RBI&I (PP,0 (PREP,3 'ON')(NP,2 (LITERAL 'THE'))
(NP2,0 (NP1,1 (ADJ,2 (QUANT,0 'FOURTH')))(NP1,0 (NP0,1 (N,6 'DAY')))))>

918 < 8 14 -46.2973: H ARIH AKL AT H $PN (S,2 (NP,2 (LITERAL 'THE')) (NP2,0
(NP1,0 (NP0,1 (N,17 'LION')))(VERB,0 (V,0 'ATE'))(NP,100
(NP,2 (LITERAL 'THE') (NP2,0 (NP1,0 (NP0,1 (N,24 'RABBIT'))))))>)

584 < 14 17 -30.6607: L ARWXH BWQR (PP,0 (PREP,6 'TO')(NP,1 (LITERAL 'A'))
(NP2,0 (NP1,0 (NNP,3 (NP0,0 (N,32 'MORNING')))(NP0,0 (N,27 'MEAL'))))))>
Stat-XFER MT Systems

• General Stat-XFER framework under development for past seven years
• Systems so far:
  – Chinese-to-English
  – French-to-English
  – Hebrew-to-English
  – Urdu-to-English
  – German-to-English
  – Hindi-to-English
  – Dutch-to-English
  – Turkish-to-English
  – Mapudungun-to-Spanish
• In progress or planned:
  – Arabic-to-English
  – Brazilian Portuguese-to-English
  – English-to-Arabic
  – Hebrew-to-Arabic
  – Czech-to-English
Syntax-based MT Resource Acquisition in Resource-rich Scenarios

- **Scenario**: Significant amounts of parallel-text at sentence-level are available
  - Parallel sentences can be word-aligned and parsed (at least on one side, ideally on both sides)
- **Goal**: Acquire both broad-coverage translation lexicons and transfer rule grammars automatically from the data
- **Syntax-based translation lexicons**:
  - Broad-coverage constituent-level translation equivalents at all levels of granularity
  - Can serve as the elementary building blocks for transfer trees constructed at runtime using the transfer rules
Syntax-driven Resource Acquisition Process

• Automatic Process for Extracting Syntax-driven Rules and Lexicons from sentence-parallel data:
  1. Word-align the parallel corpus (GIZA++)
  2. Parse the sentences independently for both languages
  3. Tree-to-tree Constituent Alignment:
      a) Run our new Constituent Aligner over the parsed sentence pairs
      b) Enhance alignments with additional Constituent Projections
  4. Extract all aligned constituents from the parallel trees
  5. Extract all derived synchronous transfer rules from the constituent-aligned parallel trees
  6. Construct a “data-base” of all extracted parallel constituents and synchronous rules with their frequencies and model them statistically (assign them relative-likelihood probabilities)
PFA Constituent Node Aligner

- Input: a bilingual pair of parsed and word-aligned sentences
- Goal: find all sub-sentential constituent alignments between the two trees which are translation equivalents of each other
- Equivalence Constraint: a pair of constituents \(<S,T>\) are considered translation equivalents if:
  - All words in yield of \(<S>\) are aligned only to words in yield of \(<T>\) (and vice-versa)
  - If \(<S>\) has a sub-constituent \(<S1>\) that is aligned to \(<T1>\), then \(<T1>\) must be a sub-constituent of \(<T>\) (and vice-versa)
- Algorithm is a bottom-up process starting from word-level, marking nodes that satisfy the constraints
• Words don’t have to align one-to-one
• Constituent labels can be different in each language
• Tree Structures can be highly divergent
Aligner uses a clever arithmetic manipulation to enforce equivalence constraints. Resulting aligned nodes are highlighted in figure 28.
Extraction of Phrases:
• Get the **yields** of the aligned nodes and add them to a phrase table tagged with **syntactic categories** on both source and target sides

• Example:
  NP # NP ::
  澳洲 # Australia
All Phrases from this tree pair:

1. IP # S :: 澳 洲 是 与 北 韩 有 邦 交 的 少 数 国 家 之 一 。 # Australia is one of the few countries that have diplomatic relations with North Korea.
2. VP # VP :: 是 与 北 韩 有 邦 交 的 少 数 国 家 之 一 # is one of the few countries that have diplomatic relations with North Korea.
3. NP # NP :: 与 北 韩 有 邦 交 的 少 数 国 家 之 一 # one of the few countries that have diplomatic relations with North Korea.
4. VP # VP :: 与 北 韩 有 邦 交 # have diplomatic relations with North Korea.
5. NP # NP :: 邦 交 # diplomatic relations.
6. NP # NP :: 北 韩 # North Korea.
7. NP # NP :: 澳 洲 # Australia.
Recent Improvements

- The **Tree-to-Tree** (T2T) method is high precision but suffers from low recall
- Alternative: **Tree-to-String** (T2S) methods (i.e. [Galley et al., 2006]) use trees on ONE side and project the nodes based on word alignments
  - High recall, but lower precision
- Recent work by Vamshi Ambati [Ambati and Lavie, 2008]: combine both methods (**T2T***) by seeding with the T2T correspondences and then adding in additional consistent projected nodes from the T2S method
  - Can be viewed as restructuring target tree to be maximally isomorphic to source tree
  - Produces richer and more accurate syntactic phrase tables that improve translation quality (versus T2T and T2S)
<table>
<thead>
<tr>
<th>TYPE</th>
<th>Total</th>
<th>TnS</th>
<th>%</th>
<th>TnT</th>
<th>%</th>
<th>O%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJP</td>
<td>600104</td>
<td>412250</td>
<td>68.6</td>
<td>176677</td>
<td>29.4</td>
<td>90.7</td>
</tr>
<tr>
<td>ADVP</td>
<td>1010307</td>
<td>696106</td>
<td>68.9</td>
<td>106532</td>
<td>10.5</td>
<td>83.1</td>
</tr>
<tr>
<td>NP</td>
<td>11204763</td>
<td>8377739</td>
<td>74.7</td>
<td>4152363</td>
<td>37.1</td>
<td>93.8</td>
</tr>
<tr>
<td>VP</td>
<td>4650093</td>
<td>2918628</td>
<td>62.7</td>
<td>238659</td>
<td>5.1</td>
<td>67.9</td>
</tr>
<tr>
<td>PP</td>
<td>3772634</td>
<td>2766654</td>
<td>73.3</td>
<td>842308</td>
<td>22.3</td>
<td>89.4</td>
</tr>
<tr>
<td>S</td>
<td>2233075</td>
<td>1506832</td>
<td>67.4</td>
<td>248281</td>
<td>11.1</td>
<td>94.5</td>
</tr>
<tr>
<td>SBAR</td>
<td>912240</td>
<td>591755</td>
<td>64.8</td>
<td>42407</td>
<td>4.6</td>
<td>91.9</td>
</tr>
<tr>
<td>SBARQ</td>
<td>19935</td>
<td>9084</td>
<td>45.5</td>
<td>7576</td>
<td>38</td>
<td>99.6</td>
</tr>
</tbody>
</table>
This is all in accordance with the principles.
• Add consistent projected nodes from source tree

• Tree Restructuring:
  – Drop links to a higher parent in the tree in favor of a lower parent
  – In case of a tie, prefer a node projected or aligned over an unaligned node
Et tout ceci dans respect des principles

T*: Restructured target tree
# Extracted Syntactic Phrases

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>The principles</td>
<td>Principes</td>
</tr>
<tr>
<td>With the principles</td>
<td>Principes</td>
</tr>
<tr>
<td>Accordance with the...</td>
<td>Respect des principes</td>
</tr>
<tr>
<td>Accordance</td>
<td>Respect</td>
</tr>
<tr>
<td>In accordance with the...</td>
<td>Dans le respect des principes</td>
</tr>
<tr>
<td>Is all in accordance with..</td>
<td>Tout ceci dans le respect...</td>
</tr>
<tr>
<td>This</td>
<td>et</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>The principles</td>
<td>Principes</td>
</tr>
<tr>
<td>With the principles</td>
<td>des Principes</td>
</tr>
<tr>
<td>Accordance with the...</td>
<td>Respect des principes</td>
</tr>
<tr>
<td>Accordance</td>
<td>Respect</td>
</tr>
<tr>
<td>In accordance with the...</td>
<td>Dans le respect des principes</td>
</tr>
<tr>
<td>Is all in accordance with..</td>
<td>Tout ceci dans le respect...</td>
</tr>
<tr>
<td>This</td>
<td>et</td>
</tr>
</tbody>
</table>

---

TnT

TnS

TnT*
Comparative Results
French-to-English

<table>
<thead>
<tr>
<th>System</th>
<th>Dev-Set BLEU</th>
<th>Test-Set BLEU</th>
<th>Test-Set METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xfer-TnS</td>
<td>26.57</td>
<td>27.02</td>
<td>57.68</td>
</tr>
<tr>
<td>Xfer-TnT</td>
<td>21.75</td>
<td>22.23</td>
<td>54.05</td>
</tr>
<tr>
<td>Xfer-TnT'</td>
<td>27.34</td>
<td>27.76</td>
<td>57.82</td>
</tr>
<tr>
<td>Xfer-Moses</td>
<td>29.54</td>
<td>30.18</td>
<td>58.13</td>
</tr>
</tbody>
</table>

- MT Experimental Setup
  - Dev Set: 600 sents, WMT 2006 data, 1 reference
  - Test Set: 2000 sents, WMT 2007 data, 1 reference
  - NO transfer rules, Stat-XFER monotonic decoder
  - SALM Language Model (430M words)
Combining Syntactic and Standard Phrase Tables

• Recent work by Greg Hanneman, Alok Parlikar and Vamshi Ambati
• Syntax-based phrase tables are still significantly lower in coverage than “standard” heuristic-based phrase extraction used in Statistical MT
• Can we combine the two approaches and obtain superior results?
• Experimenting with two main combination methods:
  – Direct Combination: Extract phrases using both approaches and then jointly score (assign MLE probabilities) them
  – Prioritized Combination: For source phrases that are syntactic – use the syntax-extracted method, for non-syntactic source phrases - take them from the “standard” extraction method
• Direct Combination appears to be slightly better so far
• Grammar builds upon syntactic phrases, decoder uses both
Recent Comparative Results
French-to-English

<table>
<thead>
<tr>
<th>Condition</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax Phrases Only</td>
<td>27.34</td>
<td>56.54</td>
</tr>
<tr>
<td>Non-syntax Phrases Only</td>
<td>30.18</td>
<td>58.35</td>
</tr>
<tr>
<td>Syntax Prioritized</td>
<td>29.61</td>
<td>58.00</td>
</tr>
<tr>
<td>Direct Combination</td>
<td>30.08</td>
<td>58.35</td>
</tr>
</tbody>
</table>

- MT Experimental Setup
  - Dev Set: 600 sents, WMT 2006 data, 1 reference
  - Test Set: 2000 sents, WMT 2007 data, 1 reference
  - NO transfer rules, Stat-XFER monotonic decoder
  - SALM Language Model (430M words)
Transfer Rule Learning

• Input: Constituent-aligned parallel trees
• Idea: Aligned nodes act as possible decomposition points of the parallel trees
  – The sub-trees of any aligned pair of nodes can be broken apart at any lower-level aligned nodes, creating an inventory of “treelet” correspondences
  – Synchronous “treelets” can be converted into synchronous rules
• Algorithm:
  – Find all possible treelet decompositions from the node aligned trees
  – “Flatten” the treelets into synchronous CFG rules
Rule Extraction Algorithm

Sub-Treelet extraction:

Extract Sub-tree segments including synchronous alignment information in the target tree. All the sub-trees and the super-tree are extracted.
Rule Extraction Algorithm

Flat Rule Creation:

Each of the treelets pairs is flattened to create a Rule in the ‘Stat-XFER Formalism’ –

Four major parts to the rule:

1. Type of the rule: Source and Target side type information
2. Constituent sequence of the synchronous flat rule
3. Alignment information of the constituents
4. Constraints in the rule (Currently not extracted)
Rule Extraction
Algorithm

Flat Rule Creation:

Sample rule:

IP::S [ NP VP .] -> [NP VP .]
(;; Alignments
(X1::Y1)
(X2::Y2)
;; Constraints
)
Rule Extraction Algorithm

Flat Rule Creation:

Sample rule:

NP::NP [VP 北 CD 有 邦交] -> [one of the CD countries that VP]

( ;; Alignments
(X1::Y7)
(X3::Y4)
)

Note:
1. Any one-to-one aligned words are elevated to Part-Of-Speech in flat rule.
2. Any non-aligned words on either source or target side remain lexicalized
All rules extracted:

**NP::NP** [VP 北 CD 有 邦交] -> [one of the CD countries that VP]

( (*score* 0.5)
;; Alignments
(X1::Y7)
(X3::Y4)
)

**IP::S** [NP VP] -> [NP VP]

( (*score* 0.5)
;; Alignments
(X1::Y1)
(X2::Y2)
)

**NP::NP** [“北韩”] -> [“North” “Korea”]

( ;Many to one alignment is a phrase
)

**VP::VP** [VC NP] -> [VBZ NP]

( (*score* 0.5)
;; Alignments
(X1::Y1)
(X2::Y2)
)

**VP::VP** [VC NP] -> [VBZ NP]

( (*score* 0.5)
;; Alignments
(X1::Y1)
(X2::Y2)
)

**NP::NP** [NR] -> [NNP]

( (*score* 0.5)
;; Alignments
(X1::Y1)
(X2::Y2)
)

**VP::VP** [北 NP VE NP] -> [VBP NP with NP]

( (*score* 0.5)
;; Alignments
(X2::Y4)
(X3::Y1)
(X4::Y2)
)
French-English System

• Large-scale broad-coverage system, developed for research experimentation
• Participated in WMT-08 and WMT-09 Evaluations
• Latest version integrates our most up-to-date processing methods:
  – French and English parsing using Berkeley Parser
  – Moses phrase tables combined with syntactic phrase tables using syntax-prioritized method
  – Very small grammar (26 rules) selected from large extracted rule set
French-English System
Data Resources

• Europarl corpus v. 4:
  – European parliamentary proceedings
  – 1.43 million sentences (36 MW)

• News Commentary corpus:
  – Editorials, columns
  – 0.06 million sentences (1 MW)

• Giga-FrEn corpus, pre-release version:
  – Crawled Canadian, European websites in various domains
  – 8.60 million sentences (191 MW)

• TOTAL:
  – about 10M sentence pairs
  – 9.57M sentence pairs after cleaning and filtering
French-English System Phrase Tables

- After complete phrase pair extraction, filtering and collapsing:
  - 424 million standard SMT phrases
  - 27 million syntactic phrases

- Combined in a syntax-prioritized combination
French-English System
Example Grammar Rules

{NP, 5256912}
NP::NP [N "de" N] -> [N N]

{NP, 5782420}
NP::NP [N ADJ] -> [ADJ N]

{VP, 2042518}
VP::VP ["ne" V "pas" VP] -> [V "not" VP]
SrcSent 1
L’extrême droite européenne est caractérisée par son racisme et son
utilisation de la question de l’immigration en tant que divergence politique.

1 0 The extreme right in Europe is characterised by its racism and use of
the immigration as a political difference.

RuleTGS: -9.9594, TransSGT: -75.2431, TransTGS: -34.7368, Frag: -0.20412,
Length: -0.0398972, Words: 24,20
SGT -1.61107 TGS -0.745873

( 0 3 "The extreme right" -188.191 "L’estreme droite"
   "(PHRS,14515871 'The extreme right')")
( 3 5 "in Europe is" -187.021 "européenne est"
   "(PHRS,113195218 'in Europe is')")
( 5 6 "characterised" -125.731 "caractérisée" "(WS,331391 'characterised')")
( 6 8 "by its" -118.707 "par son" "(PHRS,62116997 'by its')")
( 8 9 "racism" -101.507 "racisme" "(NS,300037 'racism')")
( 9 12 "and use" -176.864 "et son utilisation" "(PHRS,112444704 'and use')")
(12 17 "of the" -192.468 "de la question de l’" "(PHRS,150075588 'of the')")
(17 21 "immigration as a" -205.369 "immigration en tant que"
   "(PHRS,316257845 'immigration as a')")
(21 23 "political difference" -209.152 "divergence politique"
   "(NP,5782420 (ADJS,29478 'political') (NS,158428 'difference') )")
(23 24 "." -39.2412 "." "(PUNCTS,3074 '.)")
Current and Future Research Directions

• Automatic Transfer Rule Learning:
  – Under different scenarios:
    • From large volumes of automatically word-aligned “wild” parallel data, with parse trees on one or both sides
    • From manually word-aligned elicitation corpus
    • In the absence of morphology or POS annotated lexica
  – Compositionality and generalization
    • Granularity of constituent labels – what works best for MT?
    • Lexicalization of grammars
  – Identifying “good” rules from “bad” rules
  – Effective models for rule scoring for
    • Decoding: using scores at runtime
    • Pruning the large collections of learned rules
  – Learning Unification Constraints
Current and Future Research Directions

• Advanced Methods for Extracting and Combining Phrase Tables from Parallel Data:
  – Leveraging from both syntactic and non-syntactic extraction methods
  – Can we “syntactify” the non-syntactic phrases or apply grammar rules on them?

• Syntax-aware Word Alignment:
  – Current word alignments are naïve and unaware of syntactic information
  – Can we remove incorrect word alignments to improve the syntax-based phrase extraction?
  – Develop new syntax-aware word alignment methods
Current and Future Research Directions

- Syntax-based LMs:
  - Our syntax-based MT approach performs parsing and translation as integrated processes.
  - Our translations come out with syntax trees attached to them.
  - Add syntax-based LM features that can discriminate between good and bad trees, on both target and source sides!
Current and Future Research Directions

• Algorithms for XFER and Decoding
  – Integration and optimization of multiple features into search-based XFER parser
  – Complexity and efficiency improvements
  – Non-monotonicity issues (LM scores, unification constraints) and their consequences on search
Current and Future Research Directions

- Building Elicitation Corpora:
  - Feature Detection
  - Corpus Navigation

- Automatic Rule Refinement

- Translation for highly polysynthetic languages such as Mapudungun and Ñupiaq
Conclusions

• Stat-XFER is a promising general MT framework, suitable to a variety of MT scenarios and languages
• Provides a complete solution for building end-to-end MT systems from parallel data, akin to phrase-based SMT systems (training, tuning, runtime system)
• No open-source publically available toolkits, but extensive collaboration activities with other groups
• Complex but highly interesting set of open research issues
Questions?
Czech-to-English Translation:

MT Marathon 2009
Session Preview

Jonathan Clark
Greg Hanneman

Language Technologies Institute
Carnegie Mellon University
26 January 2009
Outline

• Stat-XFER processing pipeline
• Processed Czech–English resources
• Possible workshop tasks
  – Syntactic phrase table combination methods
  – Synchronous grammar development
    • Selection of grammar rules
    • Exploration of label granularity
    • Development of manual grammars
  – Integration of morphological analysis
Stat-XFER Data Processing

- Parsing
  - Word Alignment
    - Moses Phrase Extraction
    - Syntactic Phrase Extraction
      - Moses Phrase Table
      - Syntactic Phrase Table
      - SCFG Grammar Rules
    - Grammar Extraction
Stat-XFER Data Processing

• Corpus:
  – Project Syndicate news data: portion of CzEng corpus (84,141 sentences)
Stat-XFER Data Processing

• Parsing:
  - Czech dependency parses by TectoMT; converted to projective c-structure
  - English c-structure parses by Stanford parser
Stat-XFER Data Processing

- Word alignment:
  - GIZA++ grow-diag-final alignment done in advance on tokenized corpus
  - Alignments computed on full CzEng corpus of 8 million sentences
Stat-XFER Data Processing

- Phrase extraction:
  - Syntactic extraction by PFA node alignment algorithm, t2ts mode
  - Non-syntactic extraction with Moses package
Stat-XFER Data Processing

- Grammar extraction:
  - Using syntactic node alignments as tree decomposition points
Final Result

• Two phrase tables, with counts:

<table>
<thead>
<tr>
<th></th>
<th>NNS</th>
<th>NNS</th>
<th>rozumem</th>
<th>brains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NN</td>
<td>NN</td>
<td>rozumem</td>
<td>reason</td>
</tr>
<tr>
<td>4</td>
<td>NN</td>
<td>NN</td>
<td>rozumem</td>
<td>sense</td>
</tr>
<tr>
<td>1</td>
<td>NP</td>
<td>NP</td>
<td>rozumem</td>
<td>reason</td>
</tr>
<tr>
<td>1</td>
<td>NN</td>
<td>NN</td>
<td>rozumností</td>
<td>wisdom</td>
</tr>
<tr>
<td>1</td>
<td>JJ</td>
<td>JJ</td>
<td>rozumnou</td>
<td>sensible</td>
</tr>
<tr>
<td>1</td>
<td>ADJP</td>
<td>ADJP</td>
<td>rozumnou měrou jisté</td>
<td>reasonably certain</td>
</tr>
<tr>
<td>1</td>
<td>NP</td>
<td>NP</td>
<td>rozumnou politiku</td>
<td>sensible policy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PHR</th>
<th>PHR</th>
<th>rozumem</th>
<th>brains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem</td>
<td>reason</td>
</tr>
<tr>
<td>4</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem</td>
<td>sense</td>
</tr>
<tr>
<td>2</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem</td>
<td>sense</td>
</tr>
<tr>
<td>1</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem , a že</td>
<td>brains ; and that</td>
</tr>
<tr>
<td>1</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem , pokud</td>
<td>sense if</td>
</tr>
<tr>
<td>1</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem , pokud ne</td>
<td>sense if not</td>
</tr>
</tbody>
</table>
Final Result

• Three suffix-array language models
  – Target side of Project Syndicate corpus
  – … + more monolingual English data
  – … + target side of public CzEng corpus

• WMT tuning, development, and test sets

• = Baseline Stat-XFER system ready to analyse and expand
Outline

• Stat-XFER processing pipeline
• Processed Czech–English resources
• Possible workshop tasks
  – Syntactic phrase table combination methods
  – Synchronous grammar development
    • Selection of grammar rules
    • Exploration of label granularity
    • Development of manual grammars
  – Integration of morphological analysis
Phrase Table Combination

• Combination of non-syntactic and syntactic phrase pairs
  – Direct combination and syntax prioritization
Synchronous Grammars: Rule Selection

• Rule learning yields huge grammars
• Decoding with millions of abstract rules is intractable
• Open Question: How do we select the best grammar rules with regard to translation quality and decoding speed?
Synchronous Grammars: Label Granularity

• Rule learning assigns non-terminal and POS labels from input parse trees

• Input labels are believed appropriate…
  – For a given single language
  – According to a particular theory of grammar

• Open Question: How do we expand or collapse these labels so that they are appropriate for translating a particular language pair?
Synchronous Grammars: Czech Example

- Subject moves in English translation
- Verbs in past tense cannot be associated with modifiers in present tense

Proti odmítnutí se zitra Petr
against dismissal AUX-REFL tomorrow Peter

v práci rozhodl protestovat
of work decided to protest

“Peter decided to protest against the dismissal of work tomorrow.”

* Example from Bojar and Lopez, “Tree-based Translation,” MT Marathon Presentation 2008
Synchronous Grammars: Manual Grammar Writing

- Stat-XFER supports LFG-style unification
- Feature structures for unification can also be provided by the morphology server

\[
\text{VP::VP} : \text{[ADV VP]} \rightarrow \text{[VP ADV]}
\]

\[
(\text{(X1::Y2)}
(\text{(X2::Y1)}
(\text{(*tgsrule* 0.2)}
(\text{(*sgtrule* 0.6)}
((\text{X0 tense}) = (\text{X1 tense}))
((\text{X0 tense}) = (\text{X2 tense}))
)\)
Czech Morphology: Example

• Czech words include clitics and inflectional morphology, marking meanings such as gender and number

nerozumím
ne+rozum+ím
NEG+understand+1SG
“I do not understand”
Czech Morphology in Stat-XFER

- Stat-XFER allows external morphology server to segment and annotate words at runtime
- Ambiguous word segmentations can be encoded as a lattice
- Must segment all training data, then rebuild phrase table & language model
(Your Idea Here)

• Any ideas about applying the statistical transfer framework to Czech–English translation are welcome!