An Overview of Labelling Hiero Models

Gideon Maillette de Buy Wenniger

Institute for Logic Language and Computation
University of Amsterdam

April 26th, 2016

Outline

Ambiguity in Hiero

- single non-terminal X
  - spurious (derivational) ambiguity
    (too) many derivations
  - syntactic (linguistic) ambiguity
    overgeneration
- constraints to reduce spurious ambiguity
  - lexical anchoring, no contiguous source-side NT's, etc.

  $\Rightarrow$ undergeneration

Spurious Ambiguity

Mary kisses Peter

Mary embrasse Peter

translate: Mary loves Peter
knowing: $X \rightarrow$ aime, loves

Extracted rules:

$X \rightarrow \cdots ,$
- Mary
- Mary kisses
- $\ldots$
- Mary X
- $\ldots$
- X kisses X
- X X X

Derivations:

$X \rightarrow \cdots ,$
- X X X
- Mary, loves, Peter

$X \rightarrow \cdots ,$
- X Peter
- Mary X
- loves

$X \rightarrow \cdots ,$
- X Peter
- X X
- Mary, loves

Syntactic ambiguity

Sentence pair 1

- die Frau_1, die ein UFO gesehen_2 hat, ist nicht verrückt_3 .
- the woman_1 who has seen a UFO_2 is not crazy_3 .

$\Rightarrow$ X_1, die X_2 hat X_3 . | X_1 who has X_2 X_3 .

Sentence pair 2

- ich glaube_1, die Frau_2 hat ein UFO gesehen_3 .
- I think_1 the woman_2 has seen a UFO_3 .

$\Rightarrow$ X_1, die X_2 hat X_3 . | X_1 the X_2 has X_3 .

Syntactic ambiguity

Sentence pair 1

- die Frau_1, die ein UFO gesehen_2 hat, ist nicht verrückt_3 .
- the woman_1 who has seen a UFO_2 is not crazy_3 .

$\Rightarrow$ X_1, die X_2 hat X_3 . | X_1 who has X_2 X_3 .

Sentence pair 2

- ich glaube_1, die Frau_2 hat ein UFO gesehen_3 .
- I think_1 the woman_2 has seen a UFO_3 .

$\Rightarrow$ X_1, die X_2 hat X_3 . | X_1 the X_2 has X_3 .

Test sentence: ich glaube, die Frau hat ein UFO gesehen .

Syntactic ambiguity

Sentence pair 1

- die Frau_1, die ein UFO gesehen_2 hat, ist nicht verrückt_3 .
- the woman_1 who has seen a UFO_2 is not crazy_3 .

$\Rightarrow$ X_1, die X_2 hat X_3 . | X_1 who has X_2 X_3 .

Sentence pair 2

- ich glaube_1, die Frau_2 hat ein UFO gesehen_3 .
- I think_1 the woman_2 has seen a UFO_3 .

$\Rightarrow$ X_1, die X_2 hat X_3 . | X_1 the X_2 has X_3 .

Test sentence: ich glaube, die Frau hat ein UFO gesehen .

$\Rightarrow$ I think the woman has seen a UFO .

$\Rightarrow$ I think who has woman seen a UFO .

Undergeneration

Model expressiveness limited by

- anti-ambiguity constraints
  no contiguous NT's on source side, lexical anchoring
- decoding constraints
  max. 2 NT's on right-hand side, decoding span limit
- estimation
  limit on phrase length

Labelling to reduce ambiguity

What source of information?

- Syntactic information
  phrase-structure/dependency trees, POS tags, etc.
- Word (distribution) information
- Automatically learned clusters/categories with EM / EM-like algorithms

⇒ Question: How well do we capture context?

What perspective?

- label target/source side, and project through alignments
- label bilingual structure directly

Outline

Example of syntax resolving ambiguity

Sentence pair 1

- die Frau$_1$, die ein UFO gesehen$_2$ hat, ist nicht verrückt$_3$.
- the woman$_1$ who has seen a UFO$_2$ is not crazy$_3$.

$\Rightarrow$ NP_1, die $VBN + NP_2$ hat VP_3 . | NP_1 who has $VBN + NP_2$ VP_3 .

Sentence pair 2

- ich glaube$_1$, die Frau$_2$ hat ein UFO gesehen$_3$.
- I think$_1$ the woman$_2$ has seen a UFO$_3$.

$\Rightarrow$ $NP + VB_1$, die NN_2 hat $VBN + NP_3$. | $NP + VB_1$ the NN_2 has $VBN + NP_3$.

Test sentence: ich glaube, die Frau hat ein UFO gesehen .

$\Rightarrow$ I think the woman has seen a UFO .

$\Rightarrow$ ~~I think who has woman seen a UFO .~~

Syntax-Augmented Machine Translation (SAMT)

1. label constituent phrases only
   - Risk coverage loss in strictly syntactic systems
     "Re-structuring, Re-labeling, and Re-aligning for Syntax-Based Machine Translation" (Wang et.al, 2010)
2. add syntax **without coverage loss** w.r.t Hiero
   relaxed syntactic labels akin to Combinatorial Categorial Grammar

C : NP – the great wall

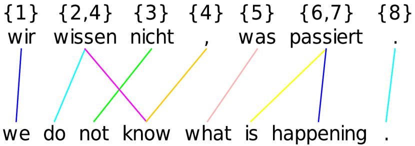
C1+C2 : NP+VB – *she+went*

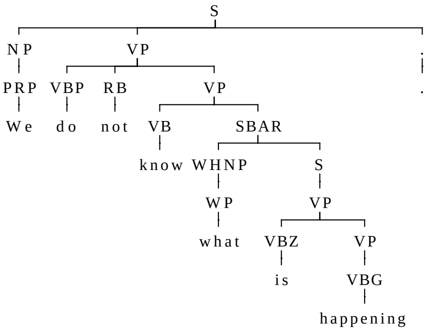
C1/C2 : NP/NN – *the great (/wall)*

C1\C2 : DT\NP – *(the\) great wall*

default : FAIL

Example

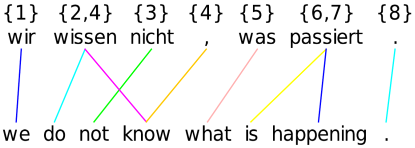


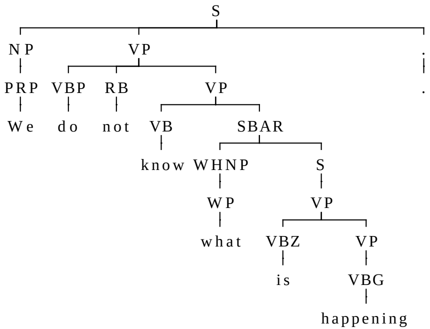


Which label for:

- we / wir (NP:PRP)
- do not know / ...
- is happening . / ...
- do ... happening . / ...
- we do not / ...

Example





Which label for:

- NP:PRP $\rightarrow$ we / wir
- VP/SBAR $\rightarrow$ do not know
- S:VP+. $\rightarrow$ is happening .
- VP+. $\rightarrow$ do ... happening .
- FAIL $\rightarrow$ we do not

Probabilistic Features

Generative probability of a rule :

- $\hat{p}(r|lhs(r))$

Phrase weights:

- $\hat{p}(r|src(r))$
- $\hat{p}(r|tgt(r))$

Phrase smoothing weights:

- $\hat{p}(r|ul(src(r)))$
- $\hat{p}(r|ul(tgt(r)))$
- $\hat{p}(ul(tgt(r))|ul(src(r)))$
- $\hat{p}(ul(src(r))|ul(tgt(r)))$

Lexical weights :

- $\hat{p}_w(tgt(r)|src(r)), \hat{p}_w(src(r)|tgt(r))$:

Results Chinese-English

Ch.-En. System \ %BLEU	Dev (MT04)	MT02	MT03	MT05	MT06	MT08	TstAvg
FULL							
Phraseb. reo=4	37.5	38.0	38.9	36.5	32.2	26.2	**34.4**
Phraseb. reo=7	40.2	40.3	41.1	38.5	34.6	27.7	**36.5**
Phraseb. reo=12	41.3*	41.0	41.8	39.4	35.2	27.9	**37.0**
Hier.	41.6*	40.9	42.5	40.3	36.5	28.7	**37.8**
SAMT	41.9*	41.0	43.0	40.6	36.5	29.2	**38.1**
TARGET-LM							
Phraseb. reo=4	35.9*	36.0	36.0	33.5	30.2	24.6	**32.1**
Phraseb. reo=7	38.3*	38.3	38.6	35.8	31.8	25.8	**34.1**
Phraseb. reo=12	39.0*	38.7	38.9	36.4	33.1	25.9	**34.6**
Hier.	38.1*	37.8	38.3	36.0	33.5	26.5	**34.4**
SAMT	39.9*	39.8	40.1	36.6	34.0	26.9	**35.5**
TARGET-LM, 10%TM							
Phraseb. reo=12	36.4*	35.8	35.3	33.5	29.9	22.9	**31.5**
Hier.	36.4*	36.5	36.3	33.8	31.5	23.9	**32.4**
SAMT	36.5*	36.1	35.8	33.7	31.2	23.8	**32.1**

Table 3.5.3: Results (% case-sensitive IBM-BLEU) for Ch-En NIST-large. Dev. scores with * indicate that the parameters of the decoder were MER-tuned for this configuration and also used in the corresponding non-marked configurations.

Results Arabic-English

Ar.-En. System \ %BLEU	Dev (MT04)	MT02	MT03	MT05	MT06	MT08	TstAvg
FULL							
Phraseb. reo=4	51.7	64.3	54.5	57.8	45.9	44.2	**53.3**
Phraseb. reo=7	51.7*	64.5	54.3	58.2	45.9	44.0	**53.4**
Phraseb. reo=9	51.7	64.3	54.4	58.3	45.9	44.0	**53.4**
Hier.	52.0*	64.4	53.5	57.5	45.5	44.1	**53.0**
SAMT	52.5*	63.9	54.2	57.5	45.5	44.9	**53.2**
TARGET-LM							
Phraseb. reo=4	49.3	61.3	51.4	53.0	42.6	40.2	**49.7**
Phraseb. reo=7	49.6*	61.5	51.9	53.2	42.8	40.1	**49.9**
Phraseb. reo=9	49.6	61.5	52.0	53.4	42.8	40.1	**50.0**
Hier.	49.1*	60.5	51.0	53.5	42.0	40.0	**49.4**
SAMT	48.3*	59.5	50.0	51.9	41.0	39.1	**48.3**
TARGET-LM, 10%TM							
Phraseb. reo=7	47.7*	59.4	50.1	51.5	40.5	37.6	**47.8**
Hier.	46.7*	58.2	48.8	50.6	39.5	37.4	**46.9**
SAMT	45.9*	57.6	48.7	50.7	40.0	37.3	**46.9**

Table 3.5.4: Results (% case-sensitive IBM-BLEU) for Ar-En NIST-large. Dev. scores with * indicate that the parameters of the decoder were MER-tuned for this configuration and also used in the corresponding non-marked configurations.

Final thoughts

- Strengths
  - as Hiero: extends phrase-based SMT
    more structure, no loss coverage
  - target-side disambiguation
    role similar to language model
- Spurious ambiguity and sparsity
  - spurious ambiguity: many labellings for same Hiero rules
  - large label set
    - ⋆ sparsity for phrase weights → smoothing essential
    - ⋆ blow-up Hiero grammar → memory problems, severe pruning at decoding, validity of best-derivation strategy ...
- and further ...
  - labelling source side (different goal and issues)
  - labelling introduces hard constraints, alternatives are soft constraints

Outline

Motivation

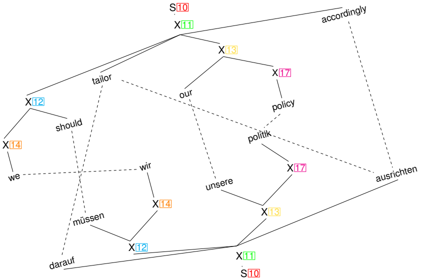
- Word alignments implicitly contain lots of reordering information
- Hiero discards almost all of this information
- Result: Hiero unable to properly model reordering at sentence level, extreme dependence on language model
- Goal: Better fulfill reordering competence promise
- Method: Effectively integrate reordering information of alignments into Hiero rules

The incoherence of translation reordering

Sentence type	Sentence contents
Source Sentence	der handlungsspielraum der beiden betroffenen regierung ist also durch das internationale recht begrenzt .
Reference	any action by the two governments concerned is therefore limited by this international law .
Hiero (Baseline)	the margin for manoeuvre of two government is concerned by the international community limited .

Hiero and Memento

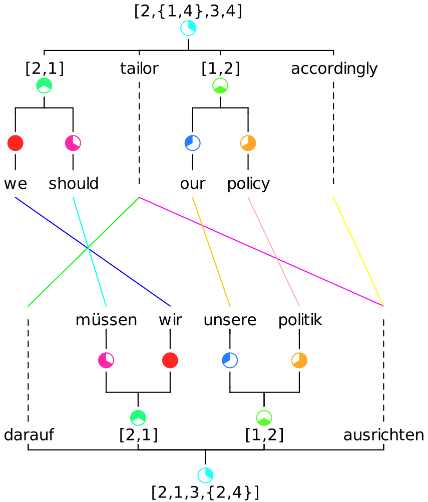
Question: what do they have in common?



Lexicalization and Language model: the words are not enough

Coherence demands (reordering) context

Vision: Hierarchical Alignment Trees (HATs)

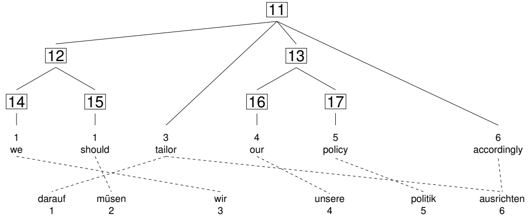


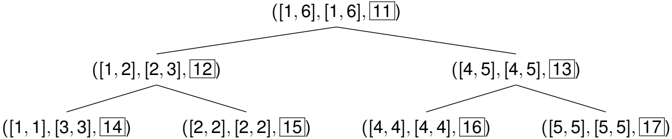
Outline

- Part 1:
  Bilingual Phrase Reordering Labels
- Part 2:
  Label Substitution Features
- Part 3:
  Experiments
- Conclusions

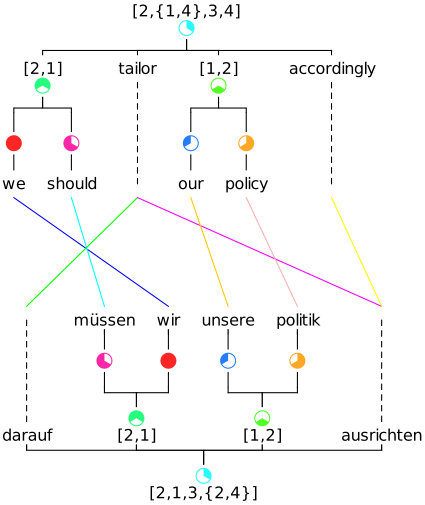
Part 1:
Bilingual Phrase Reordering Labels

NDT with Alignment structure

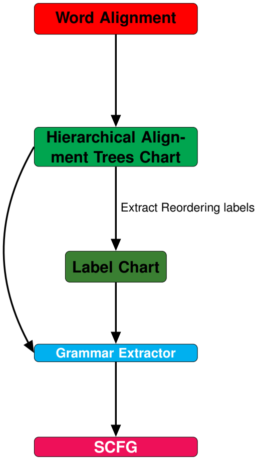




NDT with Alignment structure = HAT



Reordering Labeled Grammar Extraction

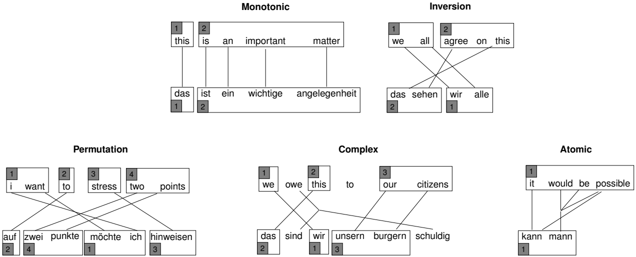


Bilingual Phrase Reordering label categories

- Phrase-Centric
- Parent-Relative

Phrase-centric reordering labels

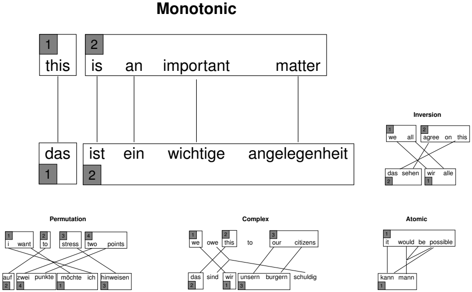
- Complexity relation between base phrase and children in *HAT* determines label
- Five cases distinguished, ordered by increasing complexity



Known labels from ITG and Phrase pair Theory

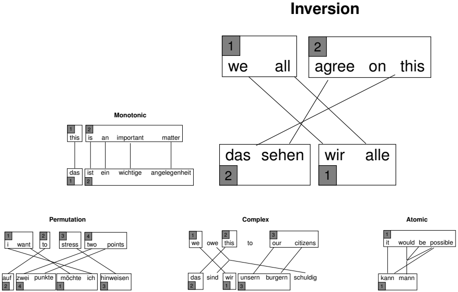
Monotonic

- *Monotonic*: If the alignment can be split into two monotonically ordered parts.



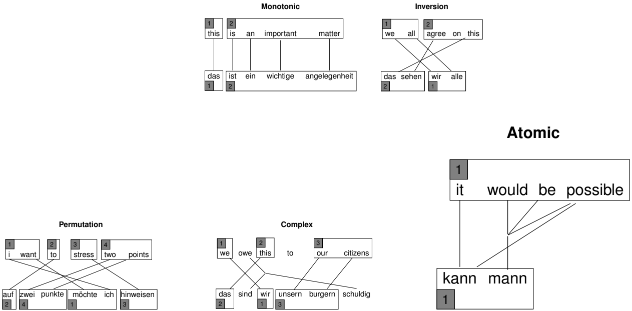
Inverted

- *Inverted*: If the alignment can be split into two inverted parts.



Atomic

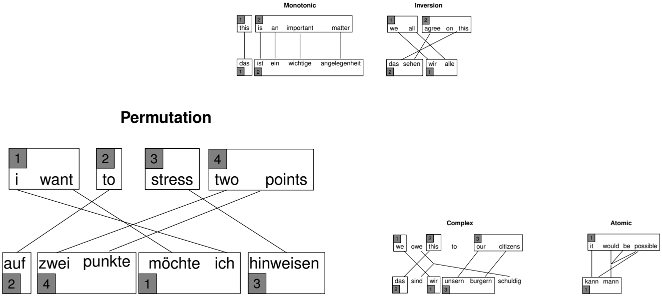
- *Atomic*: If the alignment does not allow the existence of smaller (child) phrase pairs.



New labels based on HATs

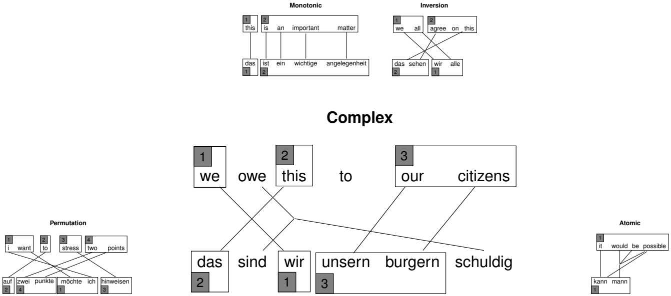
Permutation

- *Permutation*: If the alignment can be factored as a permutation of more than 3 parts.



Complex

- *Complex*: No alignment factorization as a permutation of parts, but smaller phrase pair is contained (i.e., it is composite).



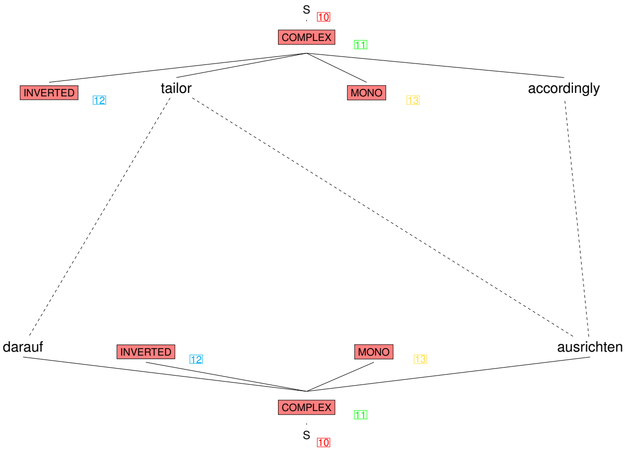
Phrase-Centric labeled derivation

S [10]

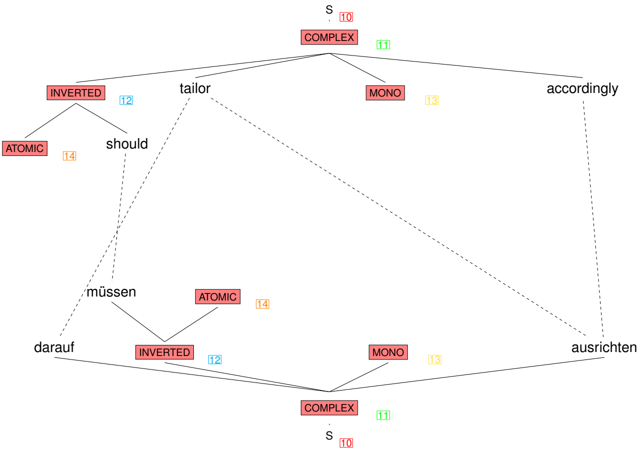
S [10]

Phrase-Centric labeled derivation

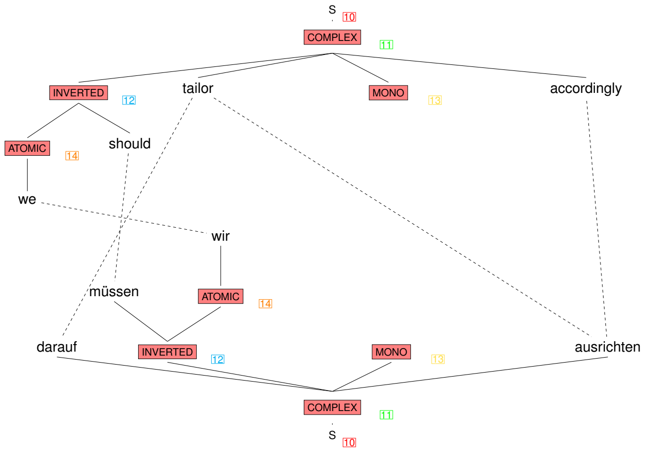
Phrase-Centric labeled derivation



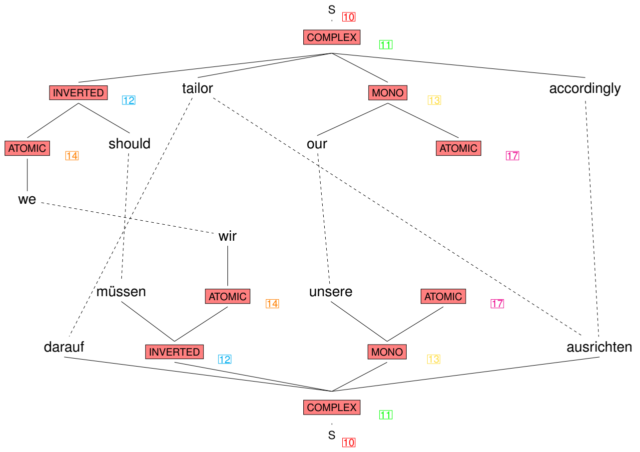
Phrase-Centric labeled derivation



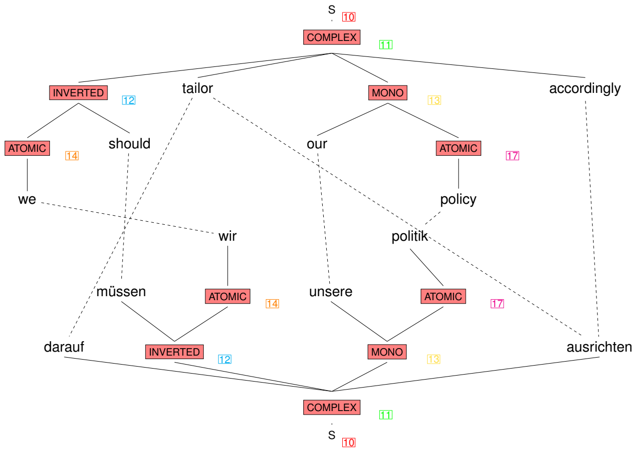
Phrase-Centric labeled derivation



Phrase-Centric labeled derivation



Phrase-Centric labeled derivation



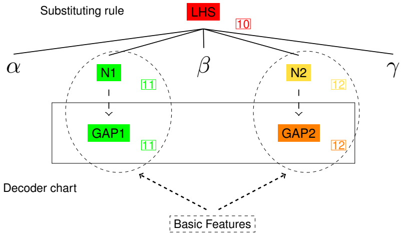
Parent-Relative reordering labels

- Describe type of reordering relative to embedding “parent” phrase
- First-order view on reordering
- (Details ommitted due to time constraints)

Part 2:
Label Substitution Features

Label substitution features

- Unique feature for every label pair $\langle L_\alpha, L_\beta \rangle$
- Marks specific LHS substitutes specific gap
- Two more coarse features:
  - Match
  - Nomatch



Part 3: Experiments

Motivating Example - After

Sentence type	**Sentence contents**
Source Sentence	der handlungsspielraum der beiden betroffenen regierung ist also durch das internationale recht begrenzt .
Reference	any action by the two governments concerned is therefore limited by this international law .
Hiero (Baseline)	the margin for manoeuvre of two government is concerned by the international community limited .
Our System	**the scope of the two governments concerned is therefore limited by international law .**

Experimental Setup

- German-English and Chinese-English language pairs
- Data properties

Language pair	dataset type	size	data origin
German-English	train	1M	Europarl
	dev	2K	WMT-07 - dev
	test	2K	WMT-07 - test
Chinese-English	train	7.34M	*MultiUn* + *Hong Kong Parallel Text*
	dev	2K	*Multiple Translation Chinese*
	test	2K	*Multiple Translation Chinese*

  - Max sentence length 40
- Language model
  - 4-gram language model
  - Kneser-Ney discounting

Experimental Setup - Evaluation

- Evaluation Metrics
  - BLEU
  - METEOR
  - Translation Error Rate (TER)
  - KENDALL-Reordering Score (KRS)
- 3 runs all experiments
- Significance Tests
  - Re-sampling test from MultEval
  - Sign test, used for KRS

Baselines

- Comparison against Hiero and SAMT baselines
- Experiments with Joshua
- Default decoding settings used

Bilingual Phrase reordering labels

Two alternative labeling schemes:

- Hiero-0th
  - Phrase-centric bilingual reordering labels
- Hiero-1st
  - Parent-relative bilingual reordering labels

Two constraint types:

- Strict constraints
- Soft constraints

Initial Results Strict Matching

System Name	DEV				TEST			
	BLEU ↑	METEOR ↑	TER ↓	KRS ↑	BLEU ↑	METEOR ↑	TER ↓	KRS ↑
	German-English							
Hiero	**27.90**	32.69	58.22	66.37	**28.39**	32.94	58.01	67.44
SAMT	27.76	32.67	58.05	**66.84**$^{\blacktriangle}$	28.32	32.88	**57.70**$^{\blacktriangle\blacktriangle}$	**67.63**
Hiero-0$^{th}_{ITG+}$	27.85	32.70	58.04$^{\blacktriangle\blacktriangle}$	66.27	28.36	32.90$^{\blacktriangledown}$	57.83$^{\blacktriangle\blacktriangle}$	67.30
Hiero-0th	27.82	**32.75**	**57.92**$^{\blacktriangle\blacktriangle}$	66.66	**28.39**	**33.03**$^{\blacktriangle\blacktriangle}$	57.75$^{\blacktriangle\blacktriangle}$	67.55
Hiero-1$^{st}_{Coarse}$	27.86	32.66	58.23	66.37	28.22$^{\blacktriangledown}$	32.90	57.93	67.47
Hiero-1st	27.74$^{\blacktriangledown}$	32.60$^{\blacktriangledown\blacktriangledown}$	58.11	66.44	28.27	32.80$^{\blacktriangledown\blacktriangledown}$	57.95	67.39
	Chinese-English							
Hiero	31.70	30.72	**61.21**	58.28	31.63	30.56	**59.28**	58.03
SAMT	**31.98**$^{\blacktriangle}$	30.81$^{\blacktriangle}$	61.83$^{\blacktriangledown\blacktriangledown}$	**60.71**$^{\blacktriangle\blacktriangle}$	31.87	30.61	59.97$^{\blacktriangledown\blacktriangledown}$	**59.94**$^{\blacktriangle\blacktriangle}$
Hiero-0$^{th}_{ITG+}$	31.54	**30.97**$^{\blacktriangle\blacktriangle}$	62.79$^{\blacktriangledown\blacktriangledown}$	59.54$^{\blacktriangle\blacktriangle}$	**31.94**$^{\blacktriangle\blacktriangle}$	**30.84**$^{\blacktriangle\blacktriangle}$	60.76$^{\blacktriangledown\blacktriangledown}$	59.45$^{\blacktriangle\blacktriangle}$
Hiero-0th	31.66	30.95$^{\blacktriangle\blacktriangle}$	62.20$^{\blacktriangledown\blacktriangledown}$	60.00$^{\blacktriangle\blacktriangle}$	31.90$^{\blacktriangle\blacktriangle}$	30.79$^{\blacktriangle\blacktriangle}$	60.11$^{\blacktriangledown\blacktriangledown}$	59.68$^{\blacktriangle\blacktriangle}$
Hiero-1$^{st}_{Coarse}$	31.64	30.75	61.37	59.48$^{\blacktriangle\blacktriangle}$	31.57	30.57	59.58$^{\blacktriangledown\blacktriangledown}$	59.13$^{\blacktriangle\blacktriangle}$
Hiero-1st	**31.74**	30.79	61.94$^{\blacktriangledown\blacktriangledown}$	**60.22**$^{\blacktriangle\blacktriangle}$	31.77	30.62	60.13$^{\blacktriangledown\blacktriangledown}$	**59.89**$^{\blacktriangle\blacktriangle}$

Main Results Soft Constraints

System Name	DEV				TEST			
	BLEU ↑	METEOR ↑	TER ↓	KRS ↑	BLEU ↑	METEOR ↑	TER ↓	KRS ↑
	German-English							
Hiero	27.90	32.69	58.22	66.37	28.39	32.94	58.01	67.44
SAMT	27.76	32.67	58.05	**66.84**▲	28.32	32.88	**57.70**▲▲	**67.63**
Hiero-0^{th}_{ITG+}-Sft	28.00▲	32.76▲▲	**57.90**▲▲	66.17	**28.48**	32.98	57.79▲▲	67.32
Hiero-0^{th}-Sft	28.01▲	32.71	57.95▲▲	66.24	28.45	32.98	57.73▲▲	67.51
Hiero-1^{st}_{Coarse}-Sft	27.94	32.69	57.91▲▲	66.26	28.45▲	32.94	57.75▲▲	67.36
Hiero-1^{st}-Sft	**28.13**▲▲	**32.80**▲▲	57.92▲▲	66.32	28.45	**33.00**▲	57.79▲▲	67.45
	Chinese-English							
Hiero	31.70	30.72	61.21	58.28	31.63	30.56	59.28	58.03
SAMT	**31.98**▲	30.81▲	61.83▼▼	**60.71**▲▲	31.87	30.61	59.97▼▼	**59.94**▲▲
Hiero-0^{th}_{ITG+}-Sft	31.88▲	30.46▼▼	**60.64**▲▲	57.82▼	31.93▲▲	30.37▼▼	**58.86**▲▲	57.60▼
Hiero-0^{th}-Sft	32.04▲▲	30.90▲▲	61.47▼▼	59.36▲▲	32.20▲▲	30.74▲▲	59.45▼	58.92▲▲
Hiero-1^{st}_{Coarse}-Sft	32.39▲▲	31.02▲▲	61.56▼▼	59.51▲▲	32.55▲▲	30.86▲▲	59.57▼▼	59.03▲▲
Hiero-1^{st}-Sft	**32.63**▲▲	**31.22**▲▲	62.00▼▼	**60.43**▲▲	**32.61**▲▲	**30.98**▲▲	60.19▼▼	**59.84**▲▲

Do we really need soft-matching?

- Best sytem strict matching (Chinese-English): 31.94 BLEU
- Best sytem soft-matching (Chinese-English): 32.61 BLEU
  - Improvement: 0.67 BLEU
- Labels are coarse (only 5 / 8 cases)
- Feature weights (Chinese-English) show strong preference matching

- Suggests soft-matching has strong merit, at least complementary (not entirely overlapping) to proper learning labels

Conclusions

- *Bilingual phrase reordering labels* improve reordering and lexical selection for Hierarchical SMT
- Using soft, not strict constraints is important to be successful
- Results also far superior to syntax-labeled translation (SAMT) for Chinese-English
- Major improvements for Chinese-English, up to $\pm$ 1 BLEU point

Outline

Source Side Labeling

- Source labels anchor rules into more (syntactic) source context
- How to assure rule labels match source labels?
  - Option 1: enforce label matching in decoder: input label chart.
  - Option 2: filter rules on the development/test set or sentence level.
- Source labels be matched only against other rule labels, as in (Chiang, 2010)

Source Side Labeling - continued

- Source labels can be thought of as adding more context, target labels as a kind of language model.
- (Mylonakis and Sima'an, 2011) illustrates how source rule labels can be matched with an input label chart, in combination with learning of rule labels.
  - Here, multiple, alternative labels per source span are allowed, with different specificity.

Questions?