What Matters Most In Morphologically Segmented SMT Models?

Mohammad Salameh

Colin Cherry

Greg Kondrak

Overview

- **Determine what steps and components of phrase-based SMT pipeline benefits the most from segmenting target language.**
- **Testing several scenarios by changing the desegmentation point in the pipeline on English-Arabic SMT system**
- **Phrases with flexible boundaries are a crucial property to a successful segmentation approach**
- **Show impact of unsegmented LMs on generation of morphologically complex words**

Segmentation/Desegmentation

Original Word	وبلعبتها	**wblEbthA/*and with her game***

Segmentation: *t* to *p*	**w+/*and***	**b+/*with***	**lEbp/*game***	**+hA/her**
	و+	ب+	لعبة	+ها

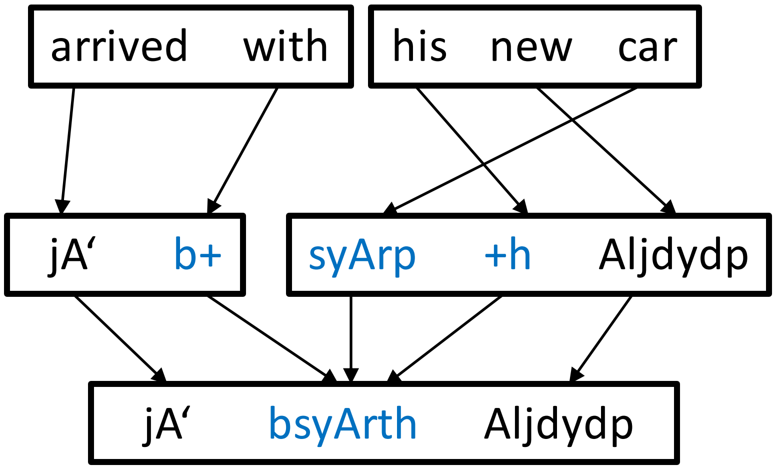
Desegmentation	وبلعبتها	**wblEbthA**

- **Morphological Segmentation is the process of segmenting words into meaningful morphemes.**
- **Desegmentation is the process of converting segmented words into their original orthographically and morphologically correct surface form**
- **Segmented vs Unsegmented vs Desegmented**

Benefits and Complications of Segmentation

English to Arabic (Morphologically Complex Language)

Benefits segmentation bring to SMT



- **Improves correspondence with morphologically simple languages**
- **Reduces data sparsity**
- **Increases expressive power by creating new lexical translations**

Complications caused by segmentation

- **Account for less context compared to word based models**
- **Less efficient statistically**
- **Introducing errors due to reversing the segmentation process at the end of the pipeline**

Measuring Segmentation Benefits

Experimental study on English to Arabic

- **Scenarios changing desegmentation point in pipeline :**
  - Before evaluation
  - Before decoding
  - Before phrase extraction
- **How these changes affect SMT component models:**
  - Alignment model, lexical weights, LM and
- **Introducing phrases with flexible boundaries**
  - suffix start: +h m$AryE fy "his projects in "
  - Prefix end: jA' b+ "arrived with"
  - Both: +hA AlAtHAd l+ "her union to"

Techniques for Morphological Segmentation/Desegmentation

Segmentation

- Penn Arabic Treebank Tokenization Scheme (El Kholy et al.[2012]) using MADA tool

Desegmentation

- Table+Rule based for Arabic (Badr et al [2008])

segmented	unsegmented	count
AbA' +km	AbAŷkm	22
AbA' +km	AbAWkm	19
DAŷqp +hm	DAŷqthm	9
kly +hA	klAhA	5

Unsegmented Baseline

- Suffers from data sparsity
- Poor correspondence
- All component models are based on words
- No desegmentation is required

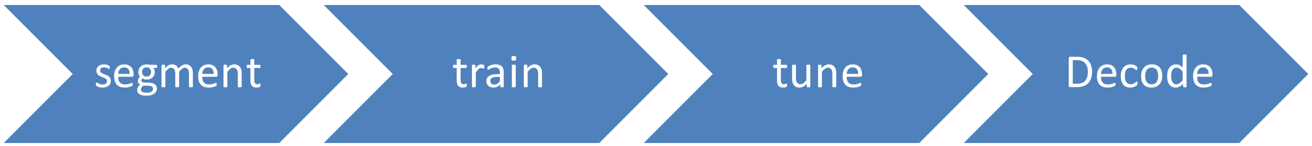
SMT components	Scenario
Desegment before	**Never Segment**
Alignment Model	**Word**
Lexical weights	**Word**
Language Model	**Word**
Tuning	**Word**
Flexible Boundaries?	**No**

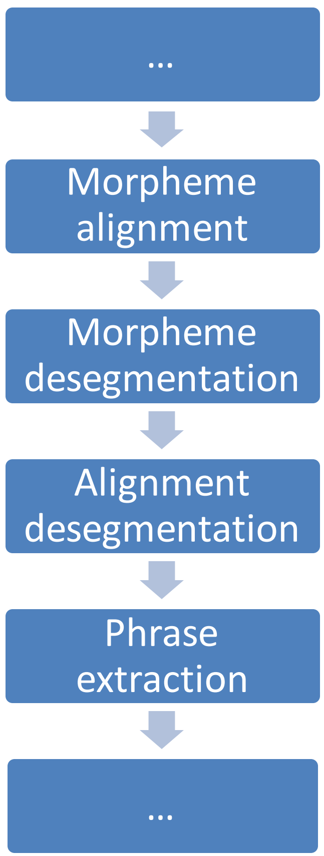
One-best Desegmentation

- Alleviates data sparsity
- improves correspondence
- All component models are based on morphemes
- LM spans shorter context
- Desegmentation is required at the end of the pipeline

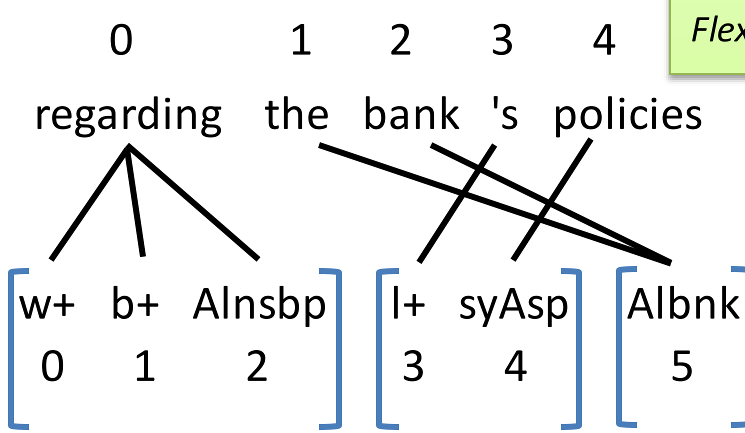
SMT components	Scenario
Desegment before	**Evaluation**
Alignment Model	**Morph**
Lexical weights	**Morph**
Language Model	**Morph**
Tuning	**Morph**
Flexible Boundaries?	**Yes**

Alignment Desegmentation

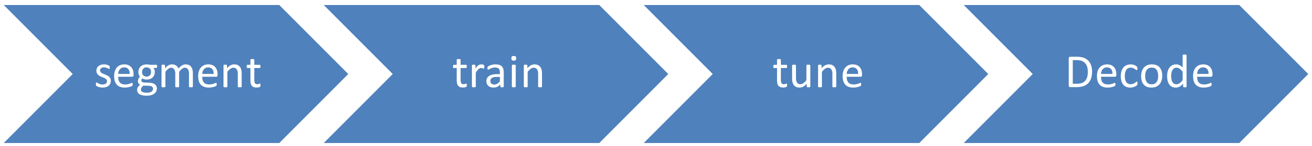


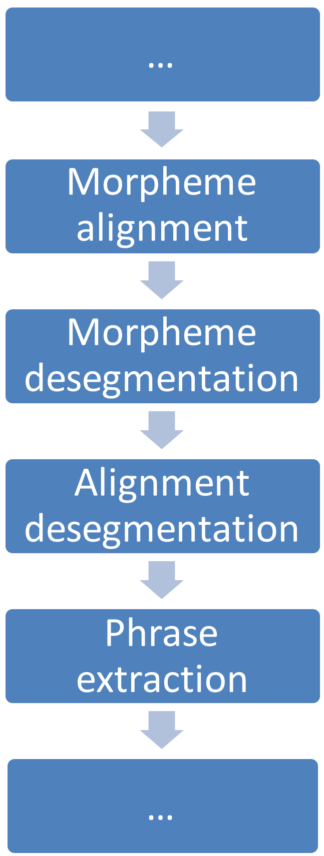


SMT components	Scenario
Desegment before	**Phrase extraction**
Alignment Model	**Morph**
Lexical weights	**Word**
Language Model	**Word**
Tuning	**Word**
Flexible Boundaries?	**No**

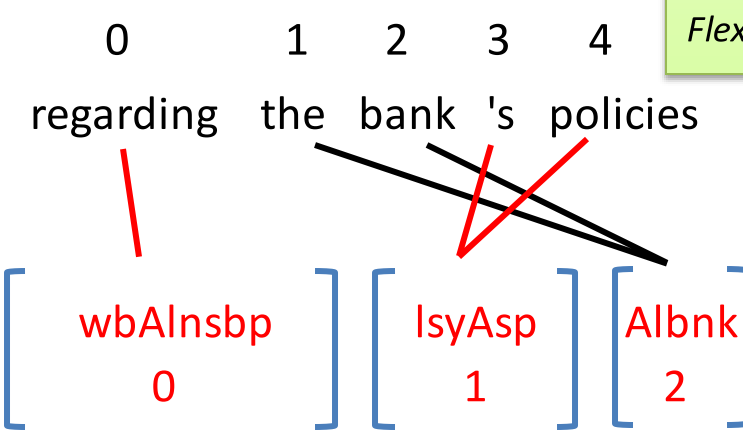


Alignment Desegmentation





SMT components	Scenario
Desegment before	**Phrase extraction**
Alignment Model	**Morph**
Lexical weights	**Word**
Language Model	**Word**
Tuning	**Word**
Flexible Boundaries?	**No**



Phrase Table Desegmentation

segment → train → tune → Decode

… → Morpheme alignment → Phrase extraction → Desegmentation

SMT components	Scenario
Desegment before	**Decoding**
Alignment Model	**Morph**
Lexical weights	**Morph**
Language Model	**Word**
Tuning	**Word**
Flexible Boundaries?	**No**

- Remove phrases with flexible boundaries from phrase table
- Desegment phrases in the phrase table
- Use word LM to score desegmented phrases

- Similar to Lyong et al. 2010

phrase with flexible boundaries

- suffix start: +h m$AryE fy "his projects in "
- Prefix end: jA' b+ "arrived with"
- Both: +hA AlAtHAd l+ "her union to"

Lattice Desegmentation (Salameh et al)

- Segment
- Train : segmented model
- Tune : using segmented reference
- Decode: generate lattice on tuning set [segmented output]
- Desegment Lattice
- Retune with added new features using unsegmented reference
- Decode on Desegmented Model

SMT components	Scenario
Desegment before	**Evaluation**
Alignment Model	**Morph**
Lexical weights	**Morph**
Language Model	**Morph+ Word**
Tuning	**Morph then Word**
Flexible Boundaries?	**Yes**

Benefits:

- gain access to a compact desegmented view of a large portion of the translation search space.
- Use features that reflect the desegmented target language
- Annotate with Unsegmented LM + Discontiguity features

Segmented LM scoring in Desegmented Models

- **Add additional LM feature that scores segmented form to :**
  - **Phrase table Desegmentation**
  - **Alignment Desegmentation**

All our problems   and conflicts

[kl   m$AklnA]   [wxlAfAtna   ]

[kl m$akl +nA]   [w+ xlAfAt +nA]

Data

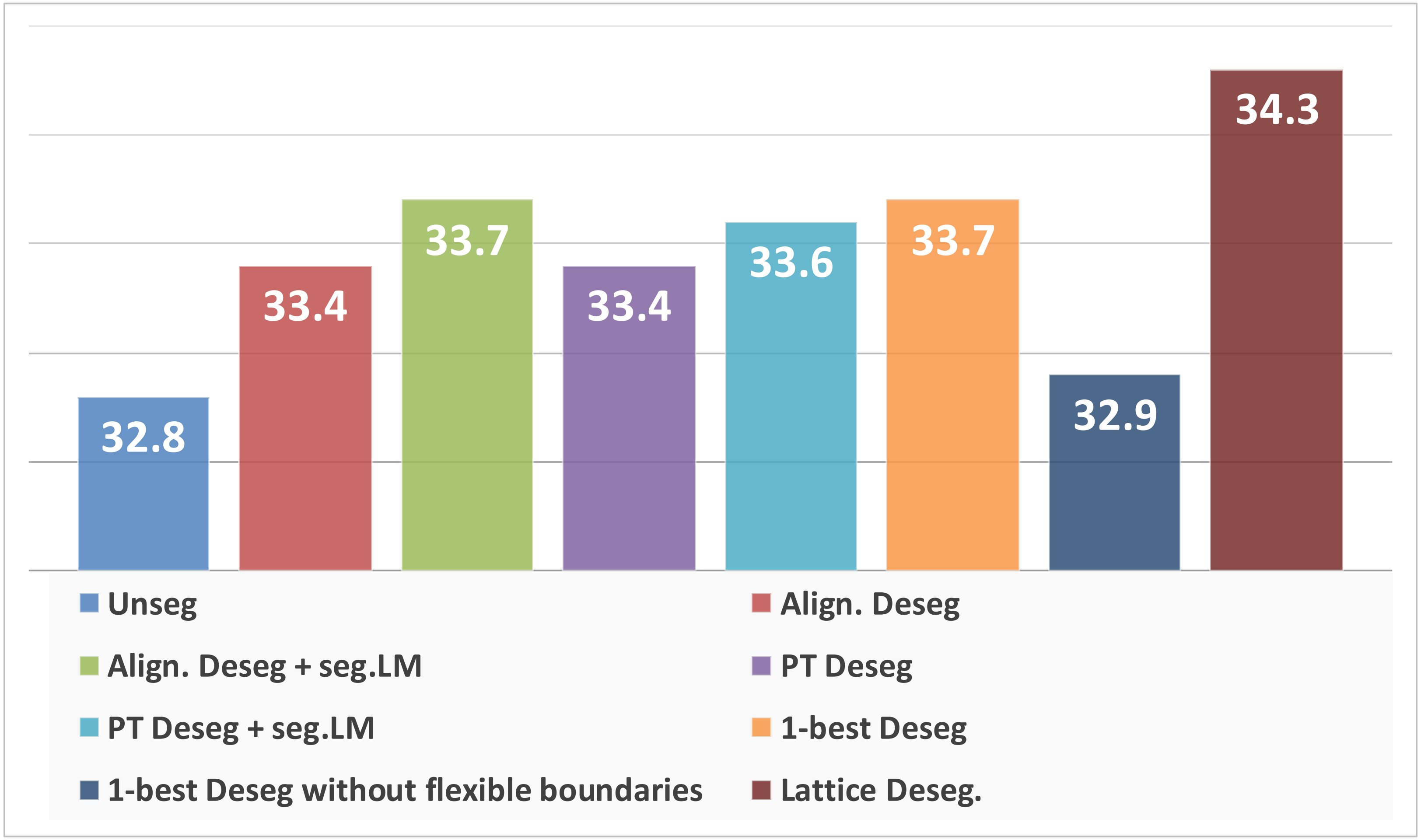
English-Arabic Data

- Train on NIST 2012 training set, excluding the UN data (1.49M sentence pairs)
- Tune on NIST 2004 (1353 pairs)
  Test on NIST 2005 (1056 pairs)
- Tune on NIST 2006 (1664 pairs)
  Test on NIST 2008 (1360 pairs)
  Test on NIST 2009 (1313 pairs)

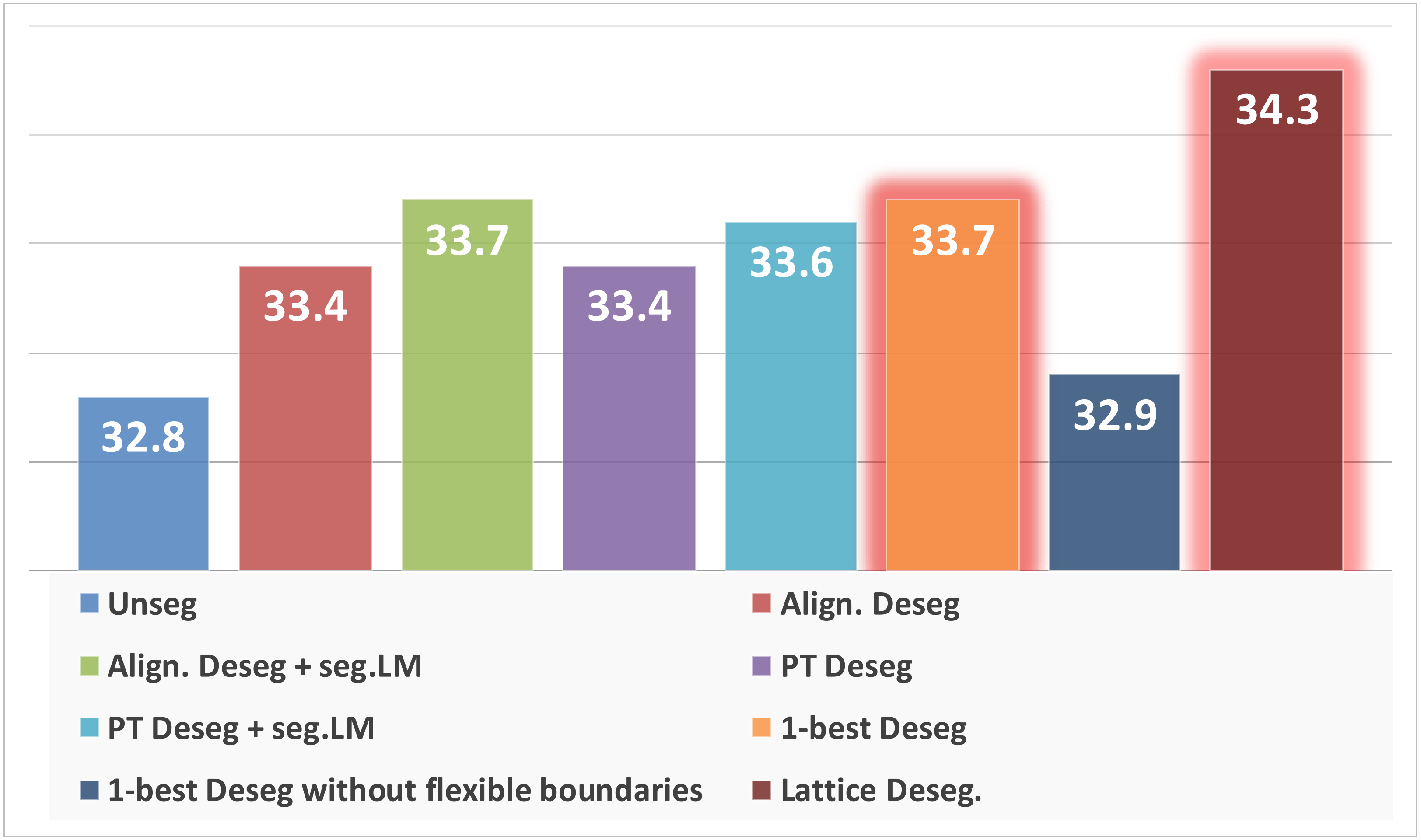
System

- **Train a 5-gram Language Model on target side using SRILM**
- **Align parallel data with GIZA++**
- **Decode using Moses**
- **Tune the decoder's log-linear model with MERT**
- **Reranking Lattice desegmented model is tuned using a batch variant of hope-fear MIRA**
- **Evaluate the system using BLEU**

Results on MT05

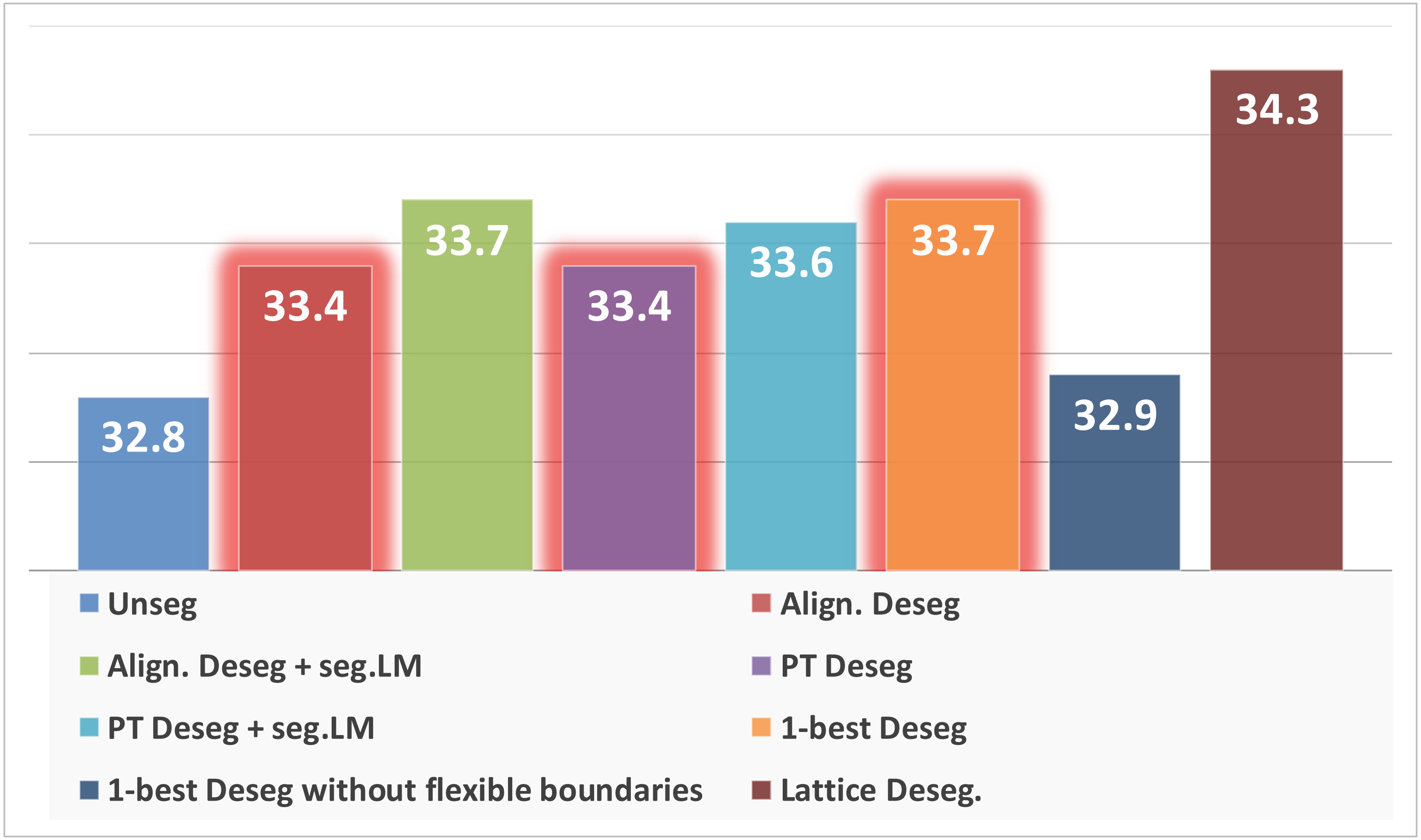


Results on MT05



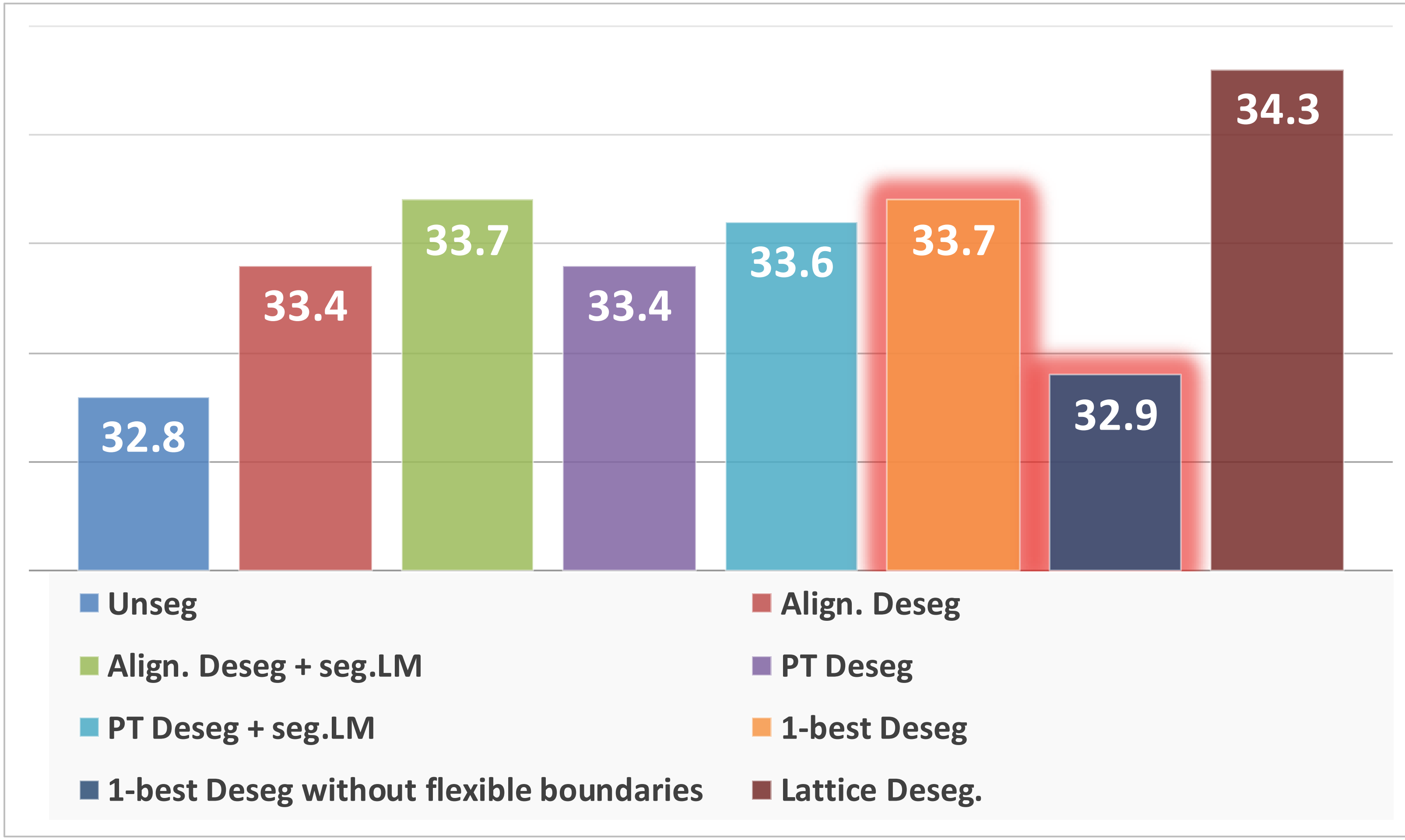
Decoder Integration: lattice desegmentation and 1-best are only systems without access to unsegmented information in the decoder

Results on MT05



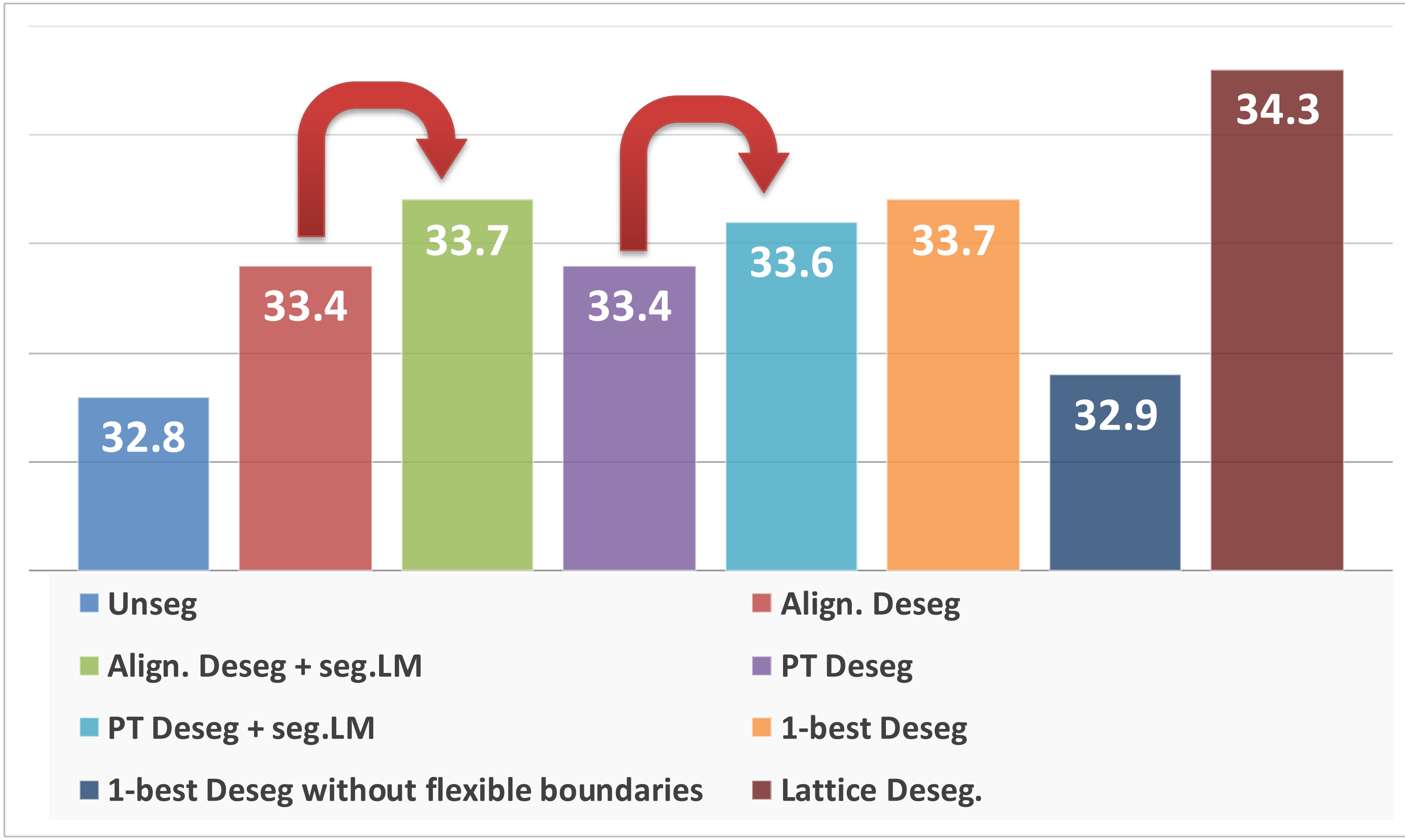
Flexible Boundaries: PT Deseg and Align Deseg. lack flexible phrase boundaries with respect to 1-best Deseg

Results on MT05



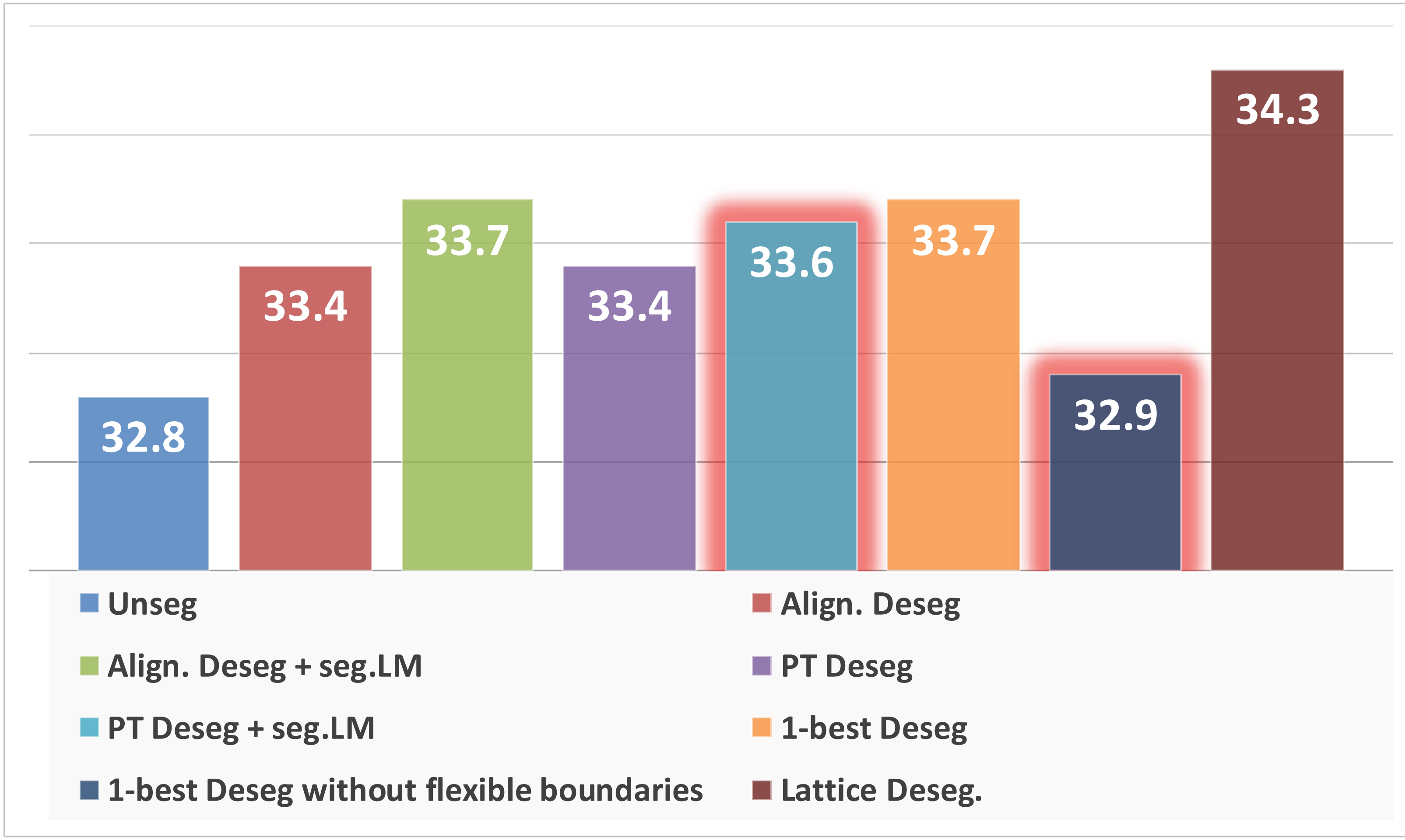
Flexible Boundaries: PT Deseg and Align Deseg. lack flexible phrase boundaries with respect to 1-best Deseg

Results on MT05



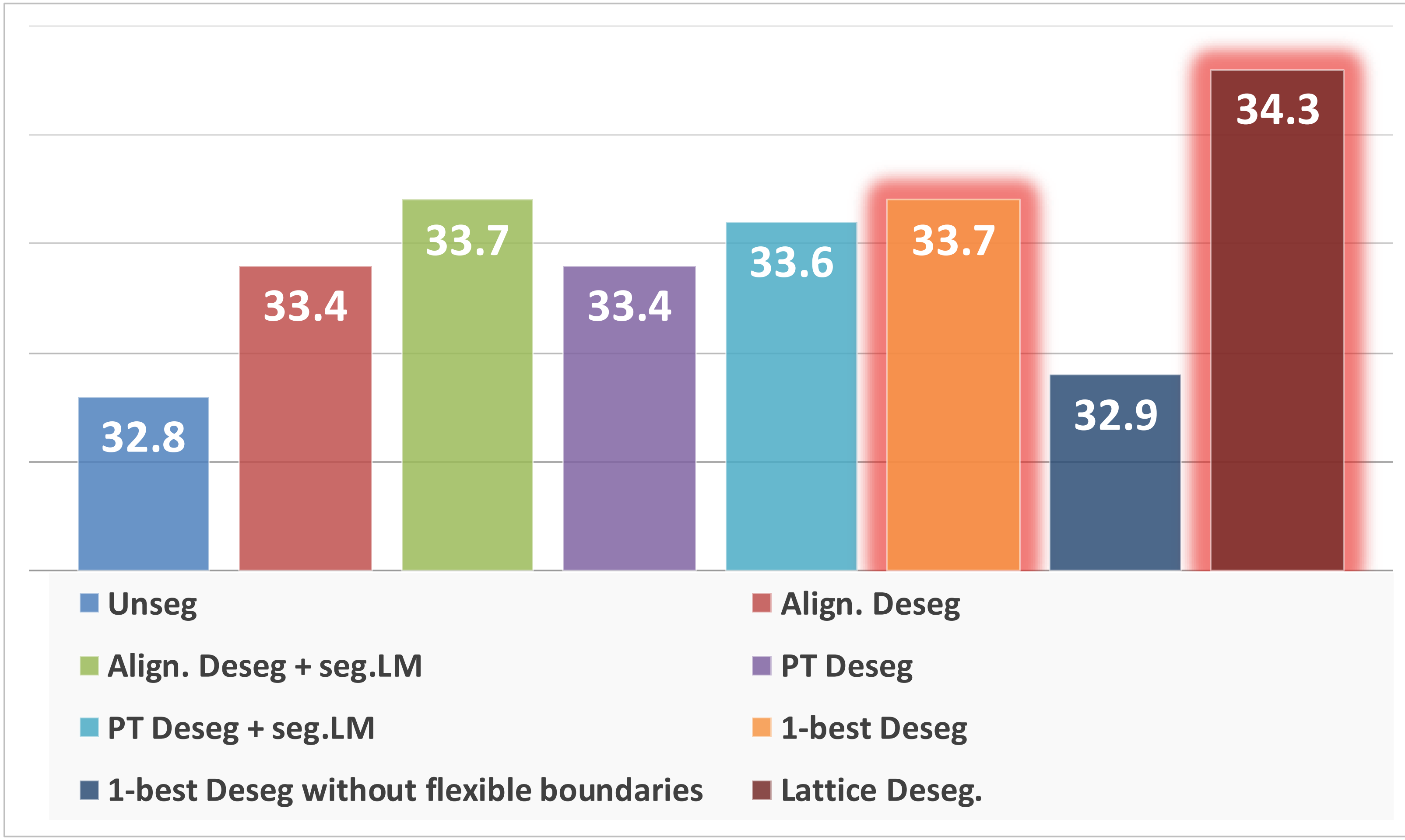
Language Models: Align Deseg and Phrase Table Deseg show consistent but small, improvements from addition of a segmented LM.

Results on MT05



Language Models: Phrase Table Deseg with segmented LM and 1-best Deseg. without flexible boundaries have exactly same output space.

Results on MT05



Language Models: main difference between 1-best Deseg. and Lattice Deseg. Is the unsegmented LM and discontiguity features.

Analysis

1. **Flexible boundaries**
   - **Constitute 12% of phrases in final output of 1-best-deseg**
   - **Novel words: 3% of the desegmented types**
     - Randomly selected 40 out of each set:
       - 64/120 violates morphological rules
     - 37/115 novel words from the reference could be constructed from morphemes
2. **Impact of *ngram* order for segmented LM**
   - **No improvement seen over 5-gram LM with 6, 7 and 8-grams**
3. **Overall affix usage**

Overall affix usage

Model	mt05	mt08	mt09
Reference	15.9	18.1	18.9
Unsegmented	12.0	12.2	12.6
Alignment Deseg.	11.6	11.0	11.8
with Segmented LM	11.7	11.2	12.0
Phrase Table Deseg.	11.3	10.1	11.2
with Segmented LM	11.6	10.5	11.4
1-best Deseg.	16.1	18.2	19.2
without flexible boundaries	14.2	14.7	15.4
Lattice Deseg.	10.0	11.5	12.2

Percentage of words in SMT output that have non-identity morphological segmentation

Overall affix usage

Model	mt05	mt08	mt09
Reference	15.9	18.1	18.9
Unsegmented	12.0	12.2	12.6
Alignment Deseg.	11.6	11.0	11.8
with Segmented LM	11.7	11.2	12.0
Phrase Table Deseg.	11.3	10.1	11.2
with Segmented LM	11.6	10.5	11.4
1-best Deseg.	16.1	18.2	19.2
without flexible boundaries	14.2	14.7	15.4
Lattice Deseg.	10.0	11.5	12.2

Percentage of words in SMT output that have non-identity morphological segmentation

Conclusion

- **Presented experimental study on translation into segmented language by creating models that apply desegmentation at different points.**
- ***Flexible boundaries* are the most important factor in improving translation in segmented models**
- **Although unsegmented LMs improve BLEU score, they hinder generation of morphologically complex words**

Thank You