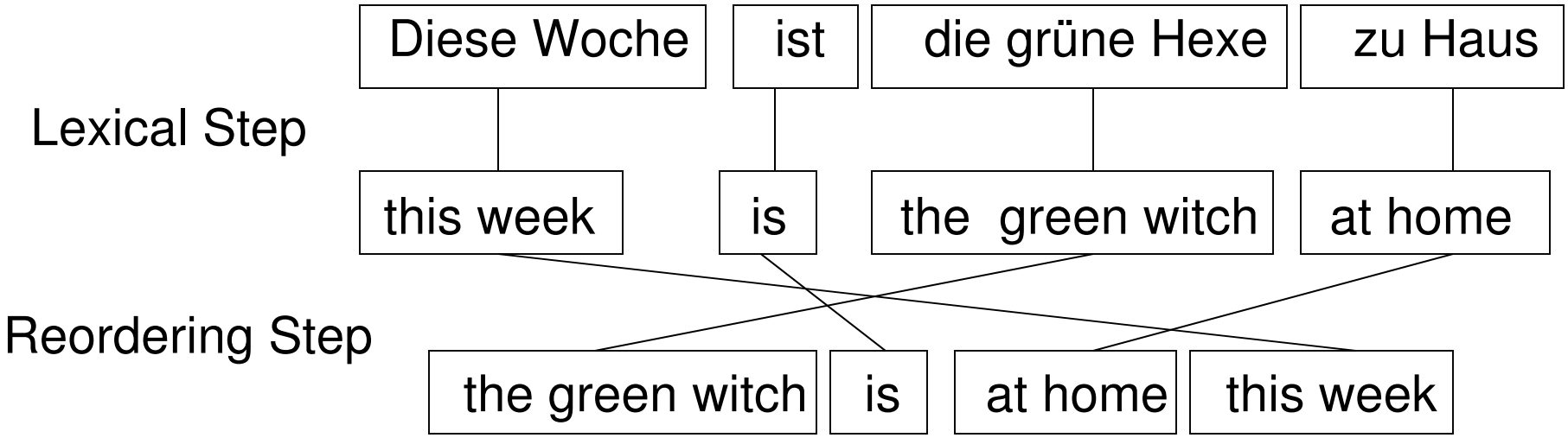
The Operation Sequence Model

Nadir Durrani
Qatar Computing Research Institute

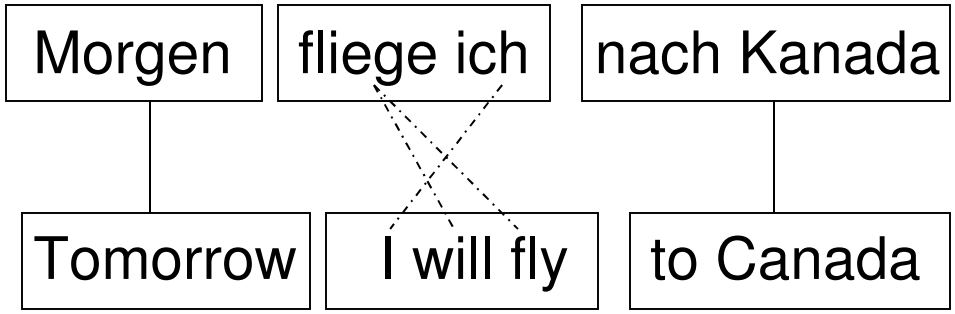
Phrase-based SMT

- State-of-the-art for many language pairs

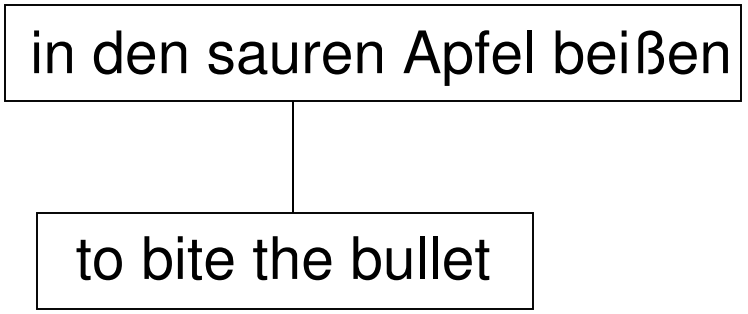


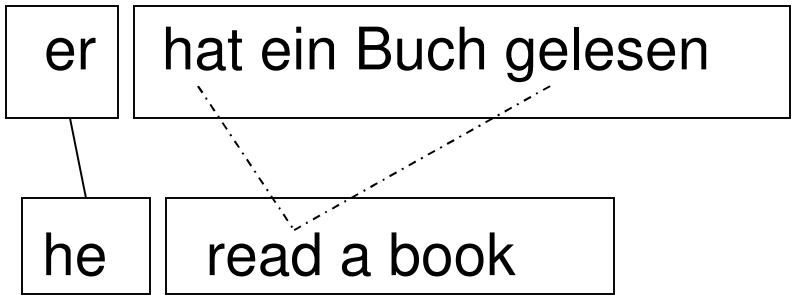
Benefits of Phrase-based SMT

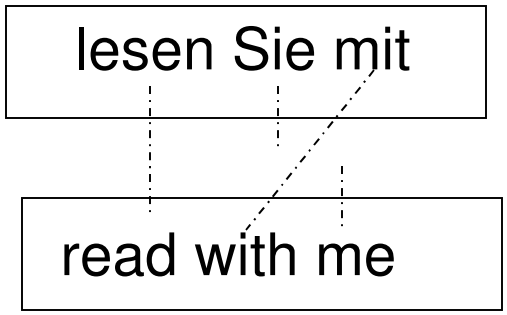
1. Local reordering



2. Idioms





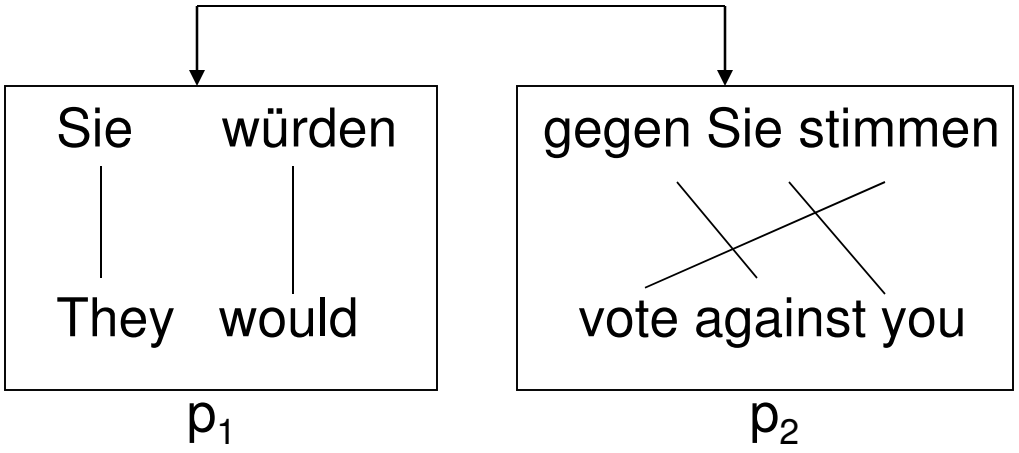


3. Discontinuities in phrases

4. Insertions and deletions

Phrase-based model: Problems

- Strong phrasal independence assumption



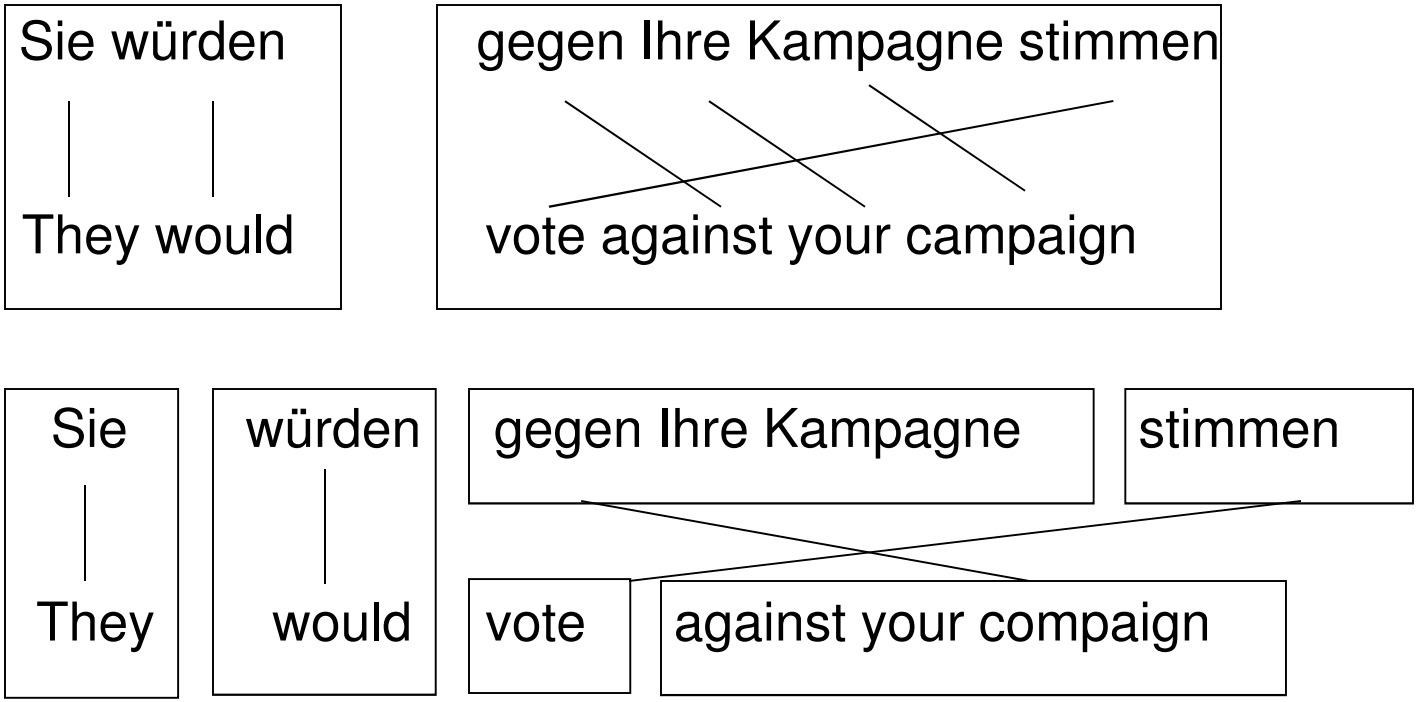
Cannot capture the dependency outside phrases

$Test_1$: Die Menschen würden gegen Sie Stimmen

$Test_2$: Die Menschen würden für die Legalisierung der Abtreibung in Kanada stimmen

Phrase-based model: Problems

- Spurious phrasal segmentation: Many possible segmentations



Phrase-based model: Problems

- Lexicalized reordering models are weak
  - Context insensitivity: Orientation only depends upon current phrase
  - Data Sparsity and Ambiguity: Short phrases have multiple orientation
    - 92.4% one word, 54% of two word Zh-EN phrases are ambiguous Li et. al (2014)
    - Most of the phrases are observed only once Cherry (2013)
  - Performs poorly with long rang reorderings
    - Heavily depends on language model
    - Most phrase-based systems use a DL=6

Motivation: Long Distance Reordering in German-to-English SMT

- Structure of main clauses in German:
  - … V2 MITTELFELD VC
- The “mittelfeld”, what’s between V2 and VC, can be arbitrarily long
- Er V2:hat ein Buch VC:gelesen → He read a book
  - hat … gelesen = read
  - Er hat gestern Nachmittag ein spannendes Buch gelesen
  - Er hat gestern Nachmittag mit seiner kleinen Tochter, die aufmerksam zugehört hat, und seinem Sohn, der lieber am Computer ein Videogame gespielt hätte, ein spannendes Buch gelesen

Motivation: Long Distance Reordering in German-to-English SMT

- Er hat ein Buch gelesen → He read a book

- Er hat gestern Nachmittag mit seiner kleinen Tochter, die aufmerksam zugehört hat, und seinem Sohn, der lieber am Computer ein Videogame gespielt haette, ein spannendes Buch gelesen

- We want a model that
  - captures ”hat … gelesen = read”
  - captures the generalization that an arbitrary amount of stuff can occur between V2:hat and VC:gelesen
  - is a simple left-to-right model

Overview

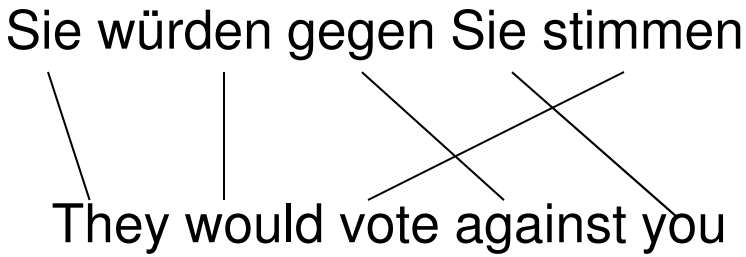
- Operation Sequence Model (OSM), a new SMT model that
  - Captures benefits of existing SMT frameworks
    - Phrase-based and N-gram-based SMT
    - Addresses their shortcomings
  - Has the ability to memorize phrase pairs
  - Robust search mechanism
  - Captures source and target information across phrasal boundaries
  - Does not have spurious phrasal segmentation ambiguity

- Unique property of OSM: Better reordering mechanism
  - Coupling of translation and reordering (like syntax-based SMT)
  - Ability to capture very long distance reordering

Introduction

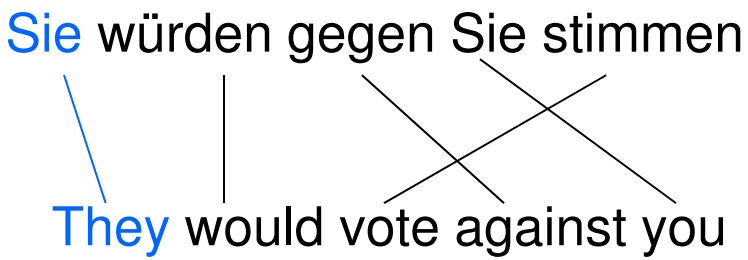
- Generation of bilingual sentence pair through a sequence of operations
- Operation: Translate or Reorder
- P(E, F, A) = Probability of the operation sequence required to generate the bilingual sentence pair

Example



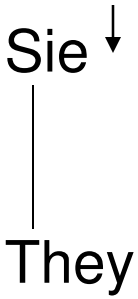
- Rules:
  - Simultaneous generate of source and target sentences
  - Generation is done in order of the target sentence
  - Reorder when source and target are in the same order

Example

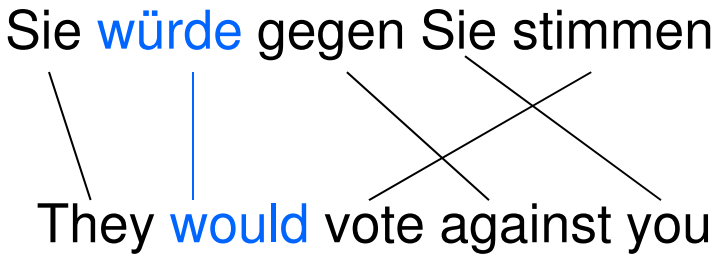


Operations

o_1: Generate (Sie – They)



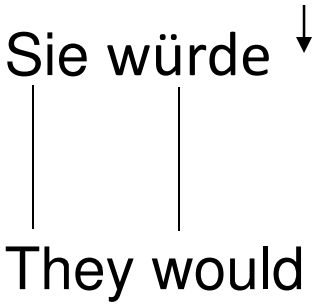
Example



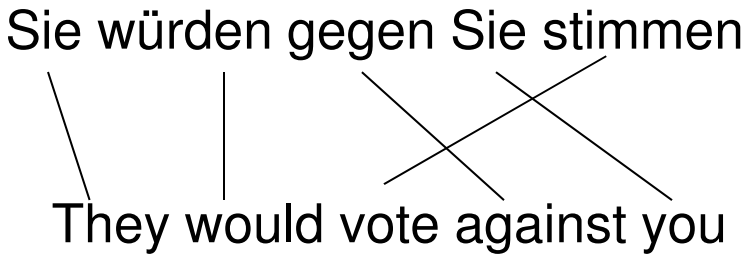
Operations

o_1 Generate (Sie, They)

o_2 Generate (würde, would)

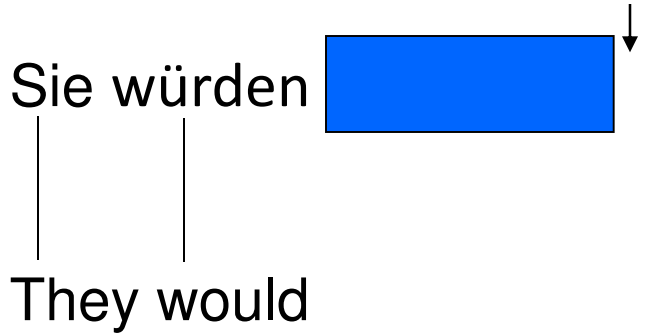


Example

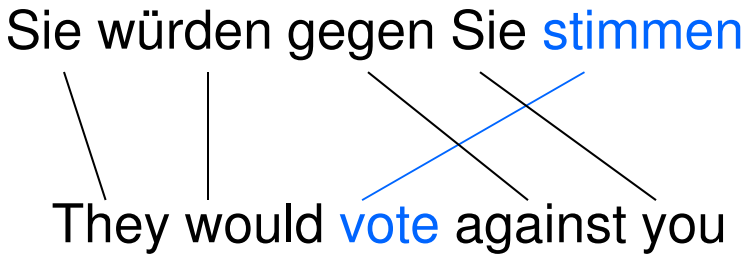


Operations

- o_1 Generate (Sie, They)
- o_2 Generate (würden, would)
- o_3 Insert Gap

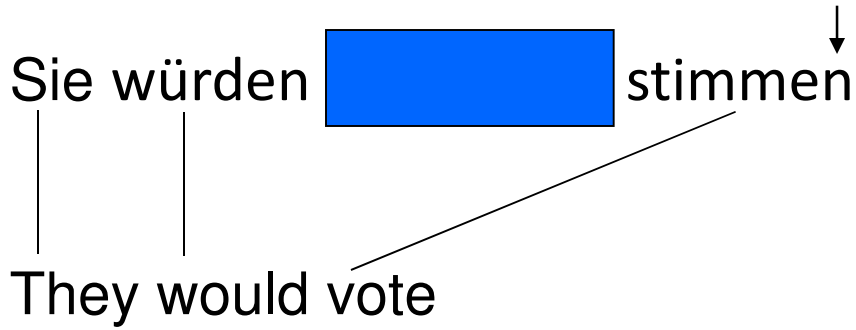


Example

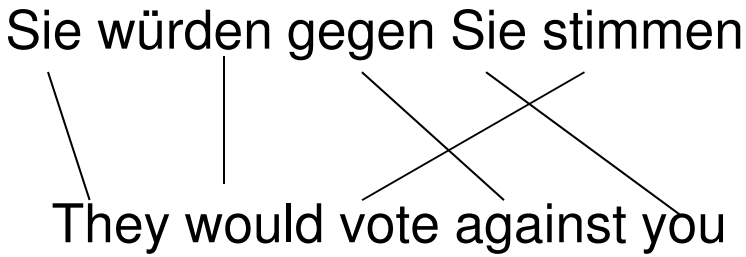


Operations

- o_1 Generate (Sie, They)
- o_2 Generate (würden, would)
- o_3 Insert Gap
- o_4 Generate (stimmen, vote)

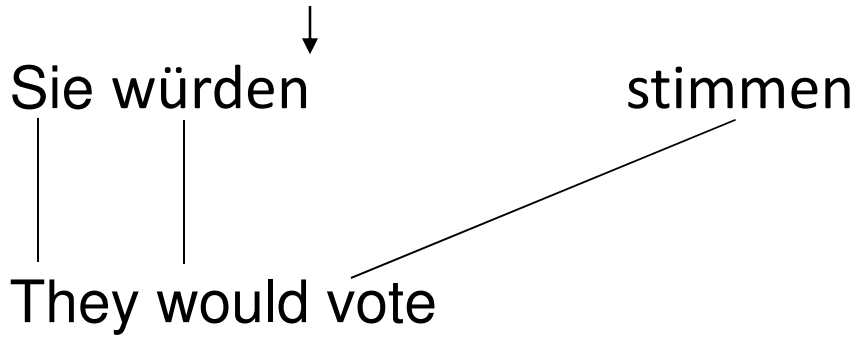


Example

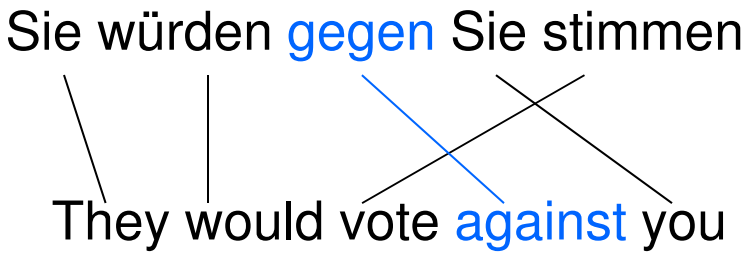


Operations

- o_1 Generate (Sie, They)
- o_2 Generate (würden, would)
- o_3 Insert Gap
- o_4 Generate (stimmen, vote)
- o_5 Jump Back (1)

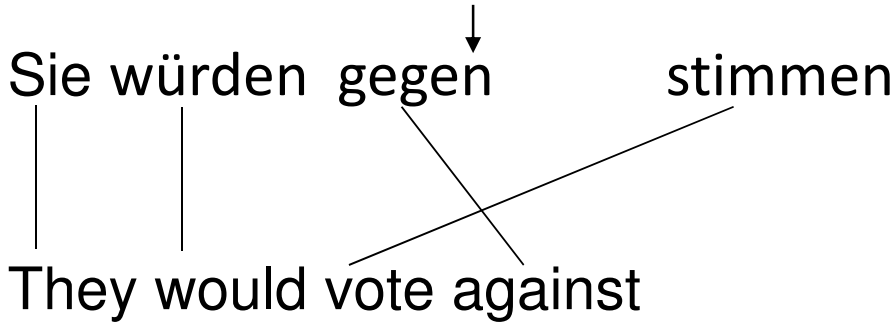


Example

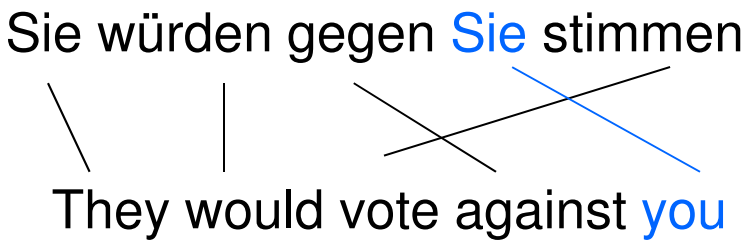


Operations

- o_1 Generate (Sie, They)
- o_2 Generate (würden, would)
- o_3 Insert Gap
- o_4 Generate (stimmen, vote)
- o_5 Jump Back (1)
- o_6 Generate (gegen, against)

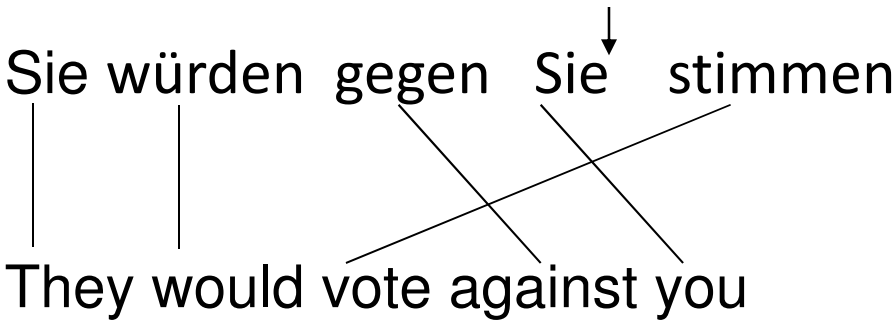


Example



Operations

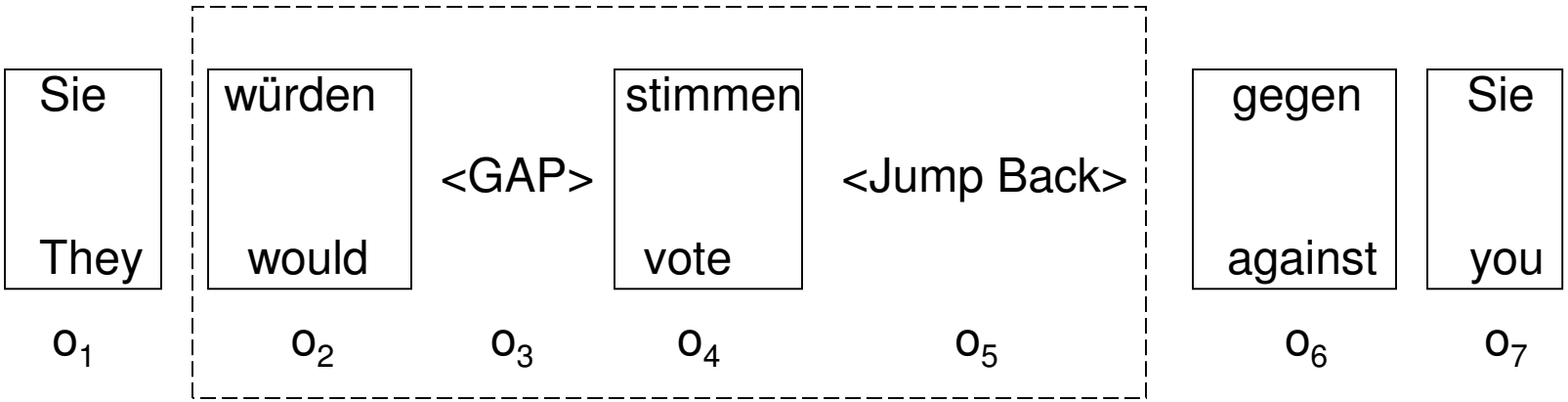
- o_1 Generate (Sie, He)
- o_2 Generate (würde, would)
- o_3 Insert Gap
- o_4 Generate (stimmen, vote)
- o_5 Jump Back (1)
- o_6 Generate (gegen, against)
- o_7 Generate (Sie, you)



Model

- Joint probability model over operation sequences

$$p_{osm}(F, E, A) = p(o_1^J) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, \ldots, o_{j-1})$$



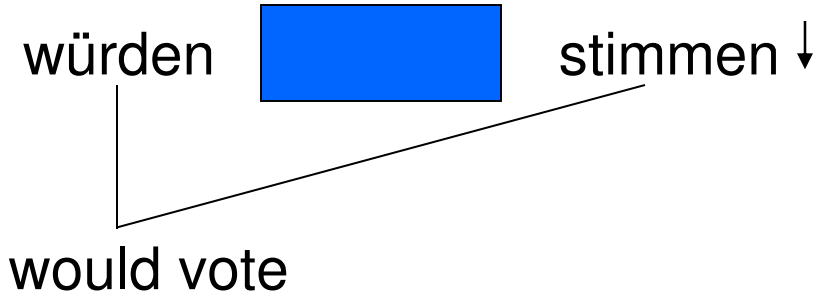
Context window: 5-gram model

Useful properties of OSM

- Capture source and target context across phrasal boundaries
  - Simultaneously generate source and target units
- Model does not have spurious ambiguity
  - Model is based on minimal translation units (MTUs)
- Better reordering mechanism
  - Uniformly handles local and non-local reorderings
  - Strong coupling of lexical generation and reordering

Example of a learned pattern

- Operations
  - Generate (würden, would)
  - Insert Gap
  - Generate (stimmen, vote)

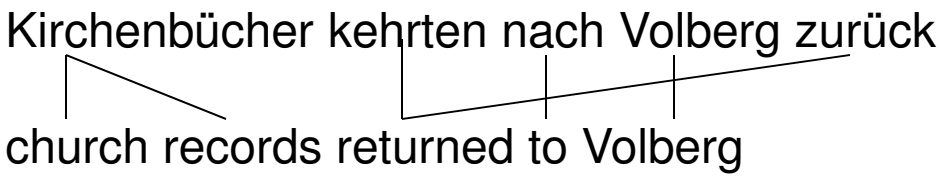


- Can generalize to
  - Die Menschen würden dafür stimmen
  - Die Menschen würden gegen meine Außenpolitik stimmen
  - Die Menschen würden für die Legalisierung der Abtreibung in Kanada stimmen

- Equivalent to hierarchical phrase “würden X stimmen – would vote X”

- Gaps can be created recursively
  - Multiple gaps can occur simultaneously

Useful properties of OSM



- Handling Discontinuities
  - kehrten [redacted] zurück – return
  - Generate (kehrten…[zurück], returned) – Insert Gap – Continue Cept
  - letzte Woche kehrten die Kinder zu ihren biologischen Eltern zurück

- Insert and Deletion Models (learning with context)
  - Source word deletion
  - Target word insertion
  - Lesen Sie mit – read with me
  - Generate (Lesen, read) → Generate Source Only (Sie) → Generate (mit, with) → Generate Target Only (me)

List of Operations

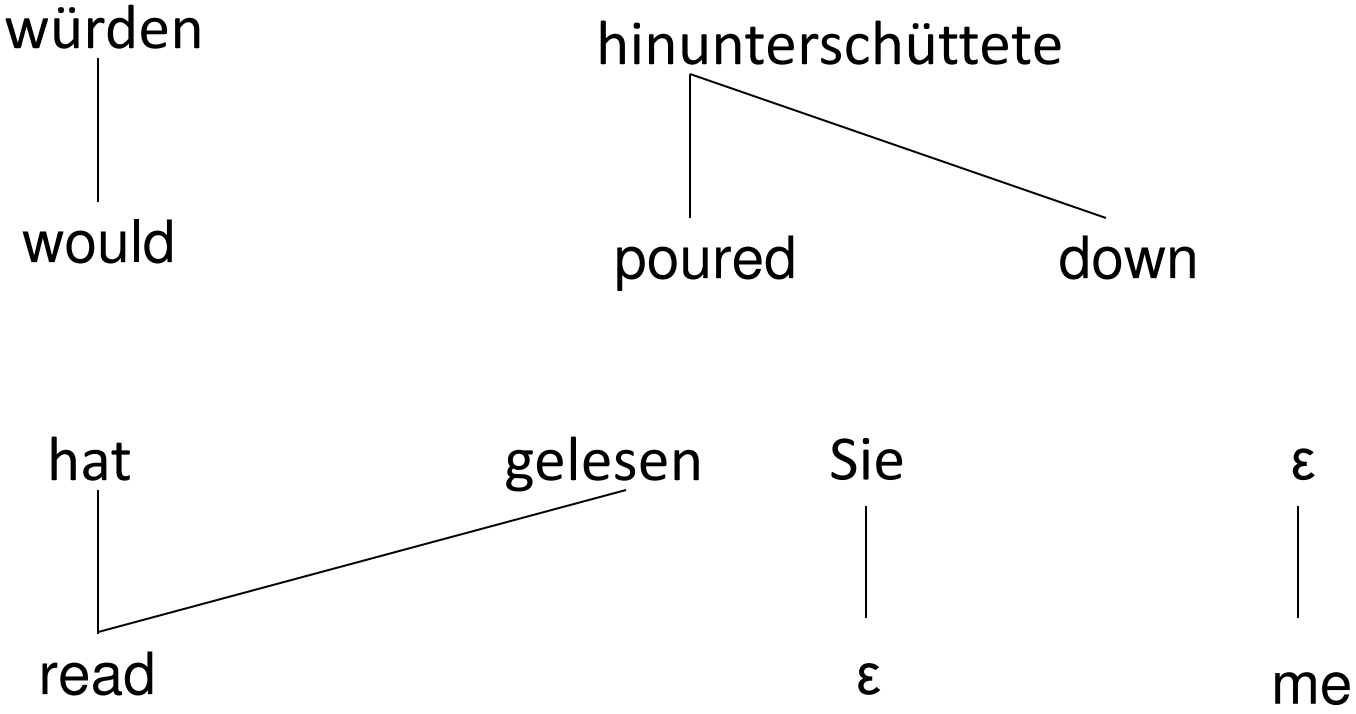
- 5 Translation Operations
  - Generate (X,Y)
  - Continue Source Cept
  - Generate Identical
  - Generate Source Only (X)
  - Generate Target Only (Y)
- 3 Reordering Operations
  - Insert Gap
  - Jump Back (N)
  - Jump Forward
- Corpus Conversion Algorithm

Operation Sequence Model (OSM) in a nutshell

- Aspects that are similar to N-gram-based SMT
  - Translation as sequential generation of a sentence pair
  - Generation is in target order
  - Sequential Markov model
  - MTU-based: Lexical operations generate MTUs
- What's new
  - Sequence of operations (as opposed to sequence of MTUs)
  - Two types of operations
    - Lexical operations
    - Reordering operations
  - Model based on minimal units, decoder based on phrases

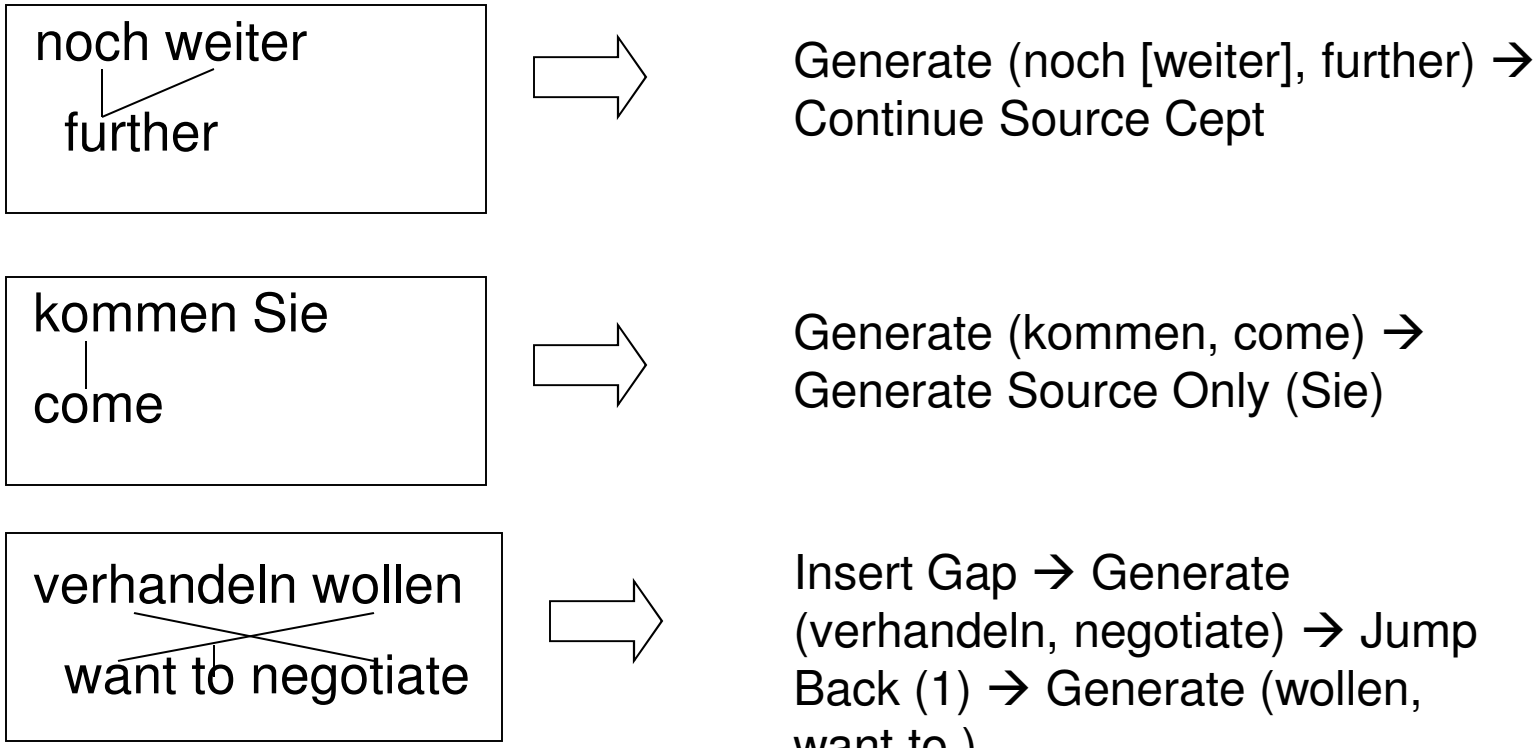
MTU = Minimal Translation Unit

- Model is based on MTUs
- Definition: MTU = connected component of the alignment bigraph



OSM learns phrases as operation sequences

- Although model is based on MTUs, we can memorize phrases



Decoders

- Cept-based decoder (or MTU-based decoding)
- Phrase-based decoder
- Moses

Decoding

- MTU-based decoding
  - extends current hypothesis by one MTU
- Phrase-based decoding
  - extends current hypothesis by one phrase
- Phrase-based decoding works better than MTU-based decoding
  - Example: Wie heißen Sie – What is your name
  - MTUs: Wie/What-is, heißen/name, Sie/your
  - Wie/What-is is very unlikely, so it will get pruned (or you need a large stack size)
  - The phrase pair: “Wie heißen Sie – What is your name” is likely
    - Makes search easier

Advantages of phrase-based decoding versus MTU-based decoding

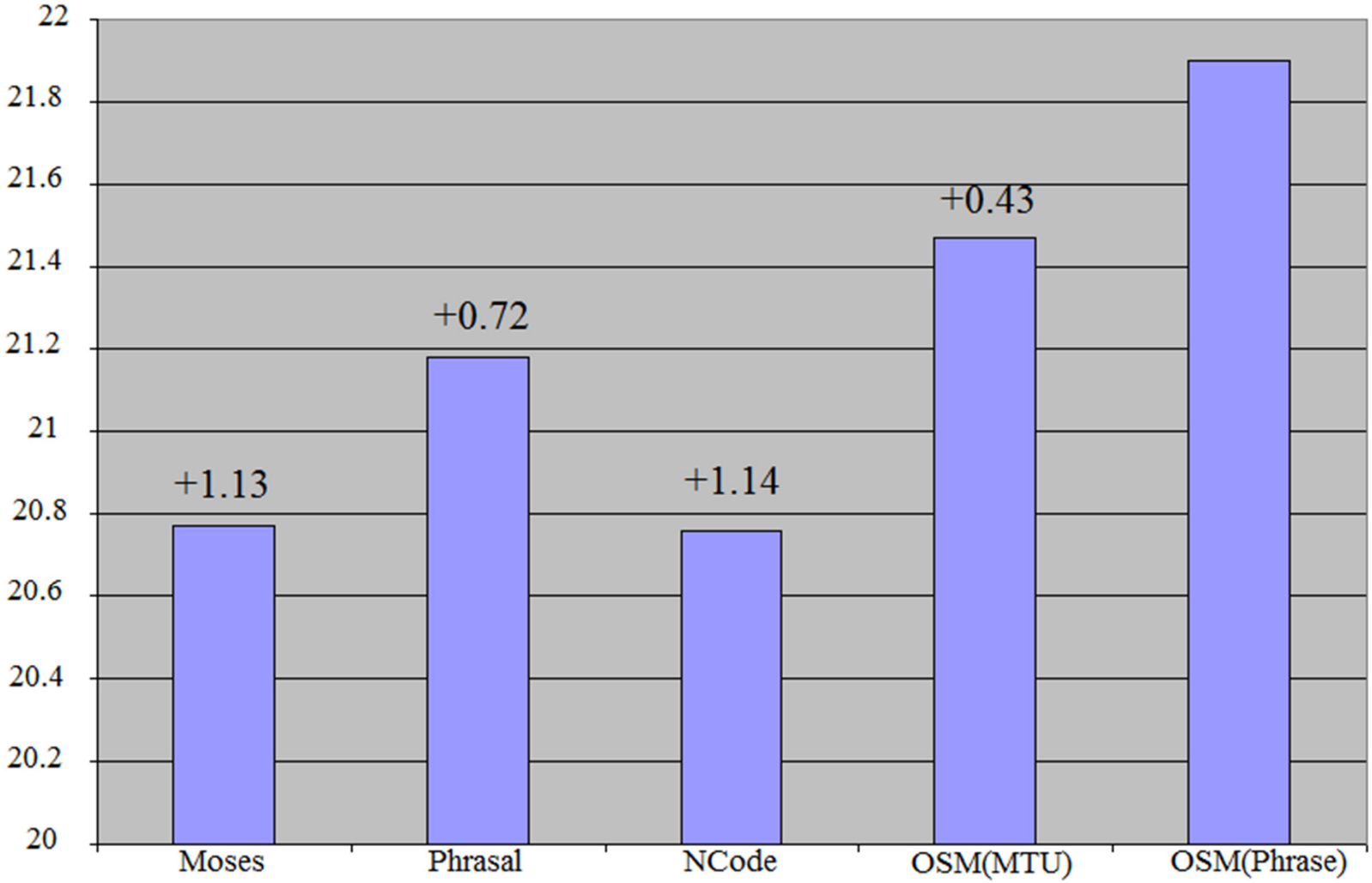
- Better future cost estimation
- Better translation coverage
- Lower beam size
- Unaligned target words
- Discontinuous target words

Baseline systems

- Moses
  - with lexicalized reordering (Koehn et. al 2005)
- Phrasal
  - with hierarchical lexicalized reordering (Galley and Manning 2008)
  - discontinuous phrases (Galley and Manning 2010)
- NCode
  - with lexicalized reordering (Crego and Yvon 2010)

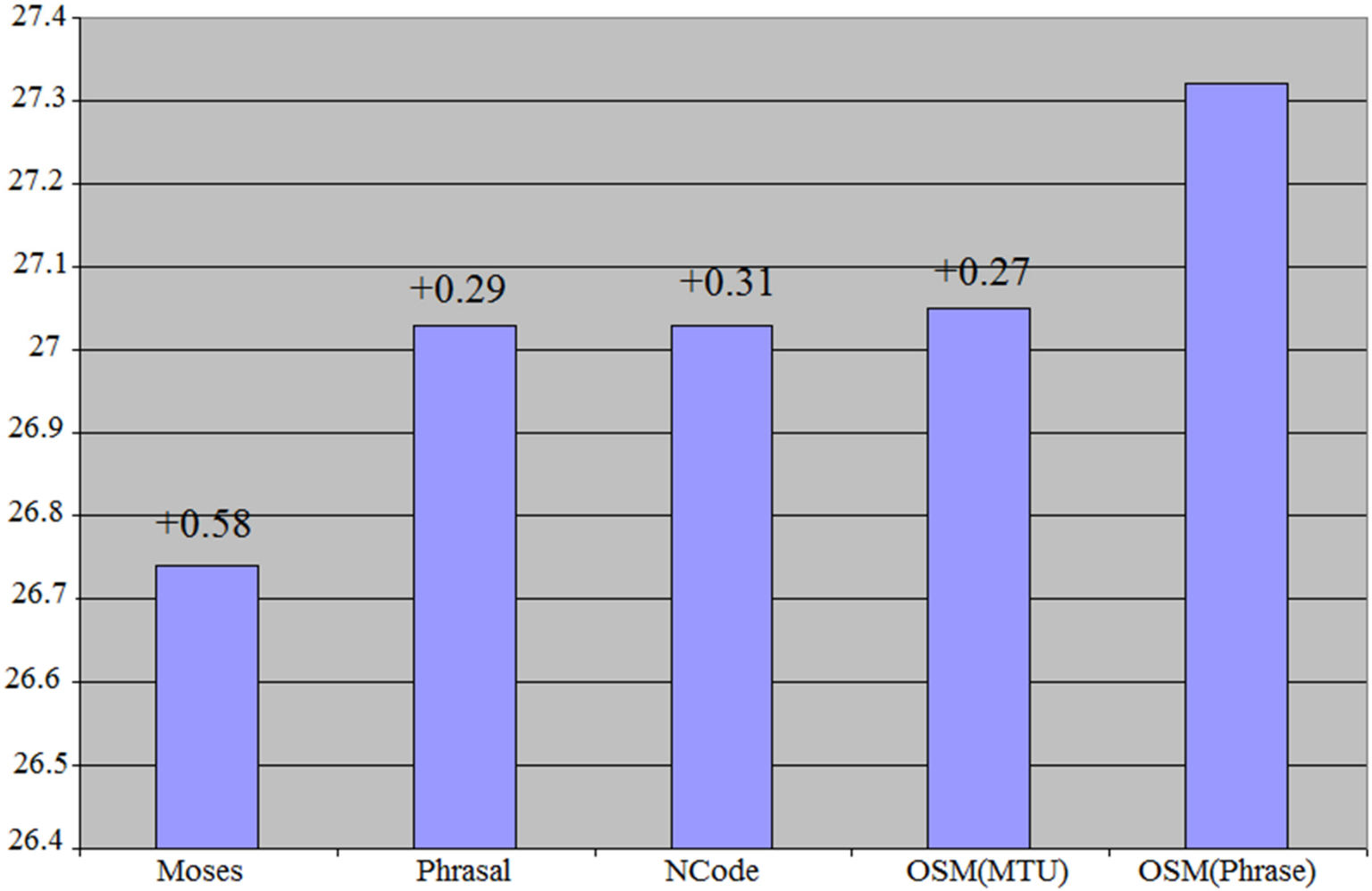
German-to-English

- Significant improvements over all baselines



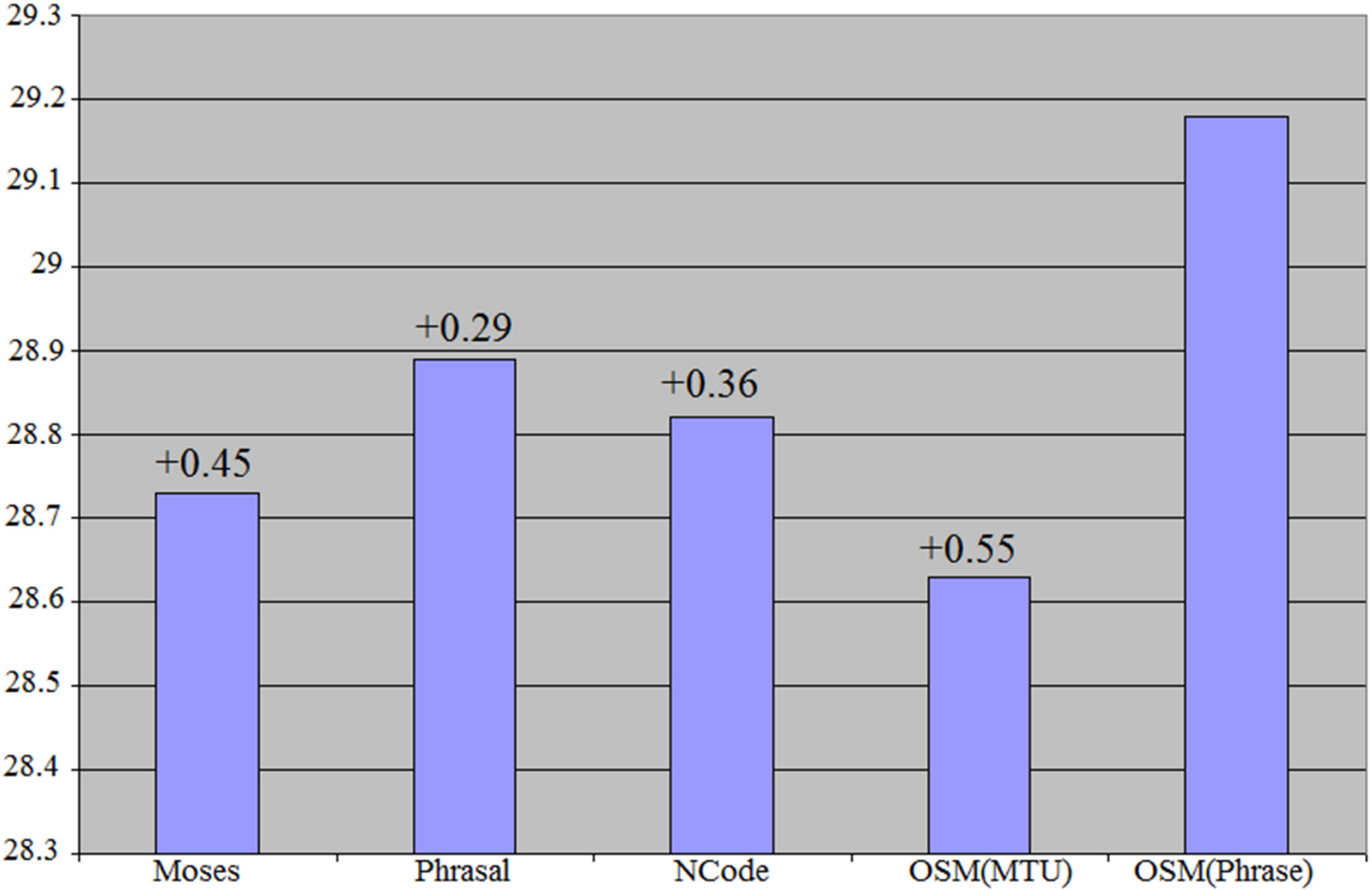
French-to-English

- Significant improvement in 11/16 cases



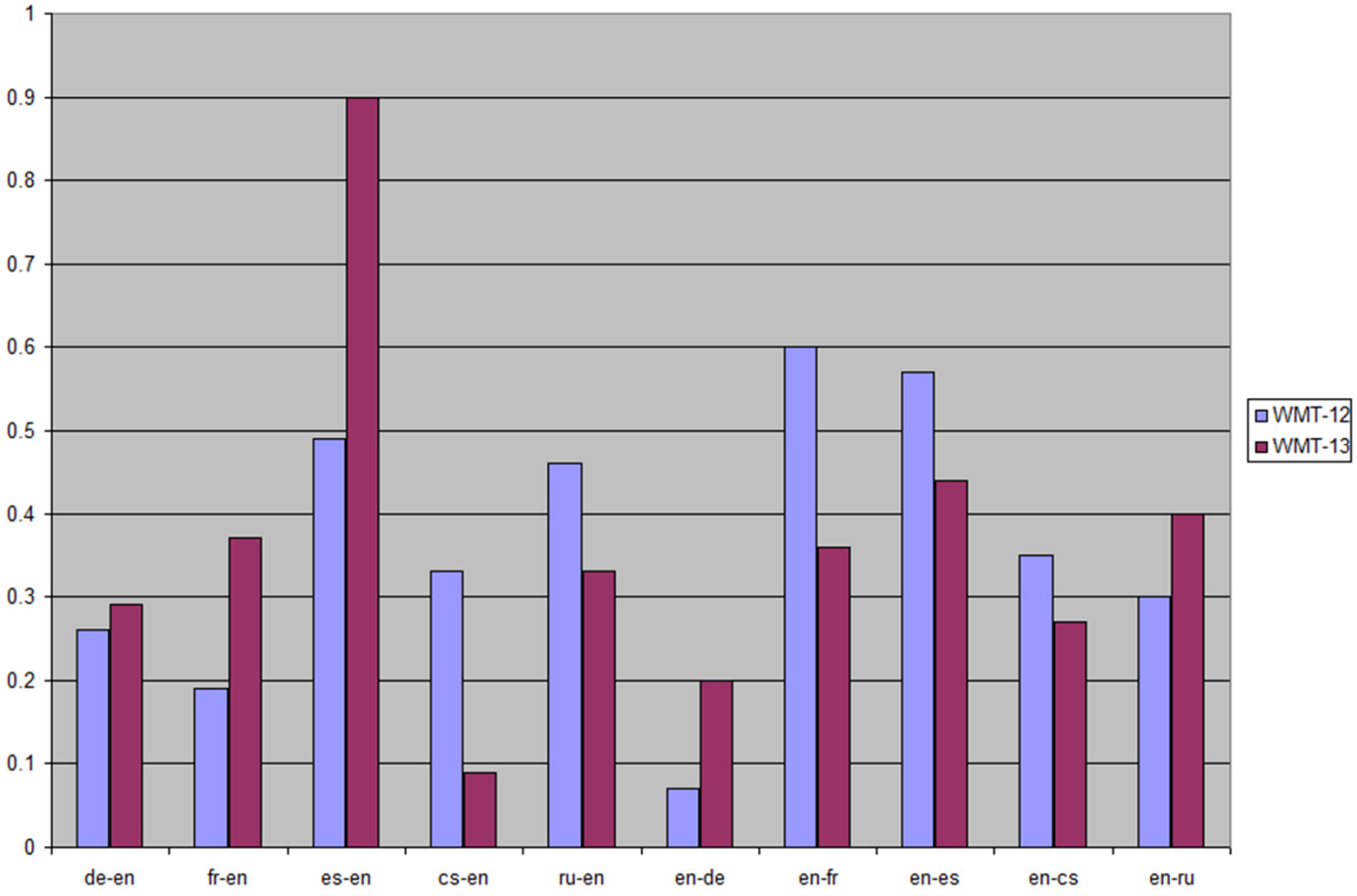
Spanish-to-English

- Significant improvement in 12/16 cases



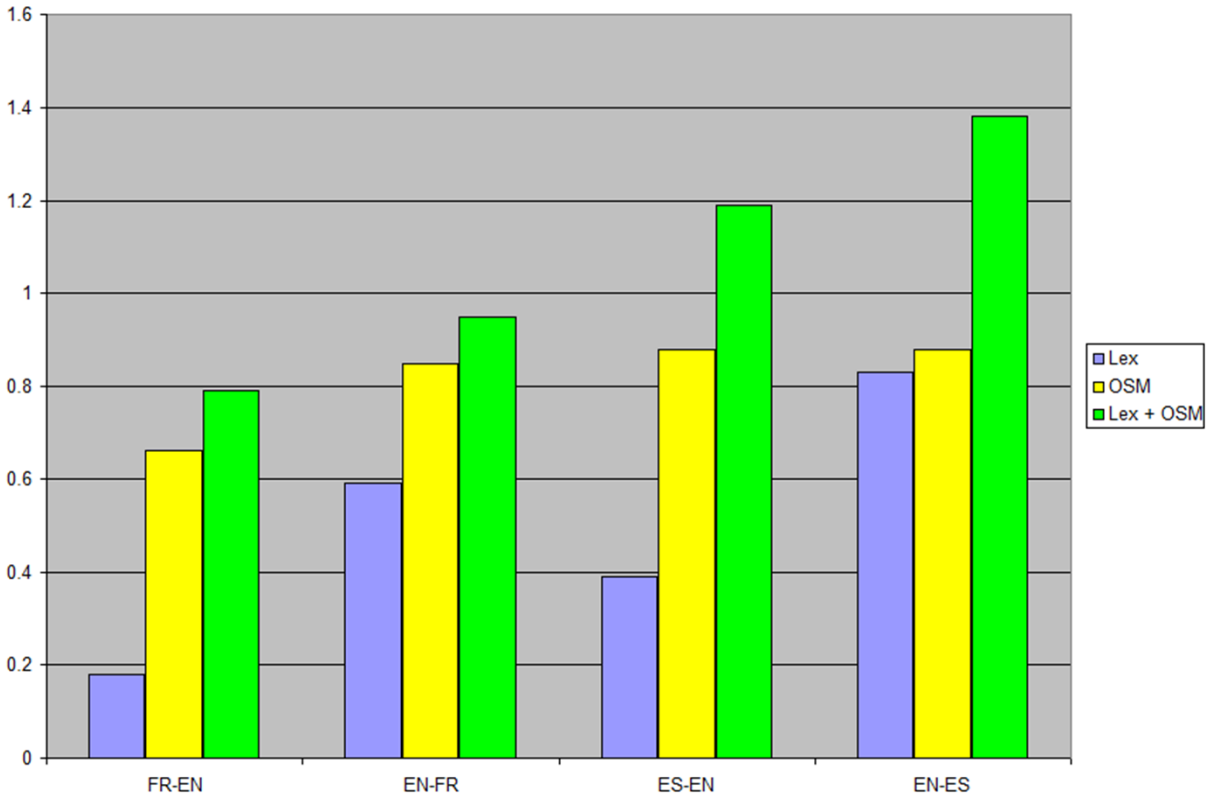
Integration into Moses (Durrani et al., ACL2013)

- Large scale evaluation (WMT-12/13)
- Average gain of +0.40 over submission quality baseline system over 10 language pairs – Significant gains in most cases



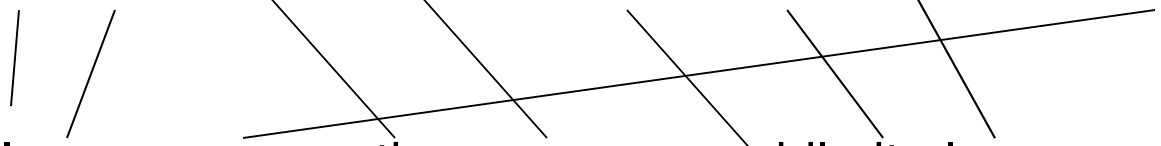
Comparison with lexicalized reordering model

- Baseline = Distortion based reordering model
  - OSM outperforms lexical reordering modeling in all four language pairs (yellow vs. purple)
  - Gains are additive (Green)



Generalized OSM Models

Ich kann die Sequenz während sie abläuft umstellen



I can rearrange the sequences while it plays

Wir können die Bücher umstellen, während er liest

We can rearrage the books while he reads

Ich kann  umstellen

PPER VMFIN  VVINF

I can rearrange

PP MD VB

Generalized OSM Models

- POS tags, Morphological Tags, Automatic Word Clusters (mkcls)

- IWSLT 2013
  - Morphological tags +1.06 for EN-DE, +0.63 for DE-EN
  - Word clusters
    - Experiments over 8 language pairs
    - Gains up to +2.02 BLEU points
  - Average OSM look-up improved from bigram to 5-gram model when using POS and automatic clusters

Summary: The OSM model …

- Integrates translation and reordering in a single generative story
- Uses bilingual context (like N-gram based SMT)
- Has an improved reordering mechanism
  - Models local and non-local dependencies uniformly
  - Takes previous translation decisions into account (like N-gram SMT)
  - Takes previous reordering decisions into account (unlike N-gram SMT)
  - Has ability to memorize lexical triggers (like phrase-based SMT)
  - Considers all possible reorderings during search

Summary: The OSM model …

- Does not have spurious phrasal segmentation (like N-gram SMT)
- Does not need ad-hoc limits during search (unlike N-gram and phrase-based)
- Supports discontinuous translation units
- Handles unaligned translation units through built-in insertion and deletion models

Conclusion

- OSM = union of N-gram-based SMT and phrase-based SMT
- Phrase-based decoding+OSM is an effective combination of an MTU-based model and phrase-based-search
- Can be used as a feature in any left-to-right decoding mechanism
- Statistically significant improvements over baseline systems

Conclusion

- Model has been extended to use in tandem with factored-based models
  - Wider context
  - Better generalization
  - Improved translation and reordering decisions
- Operation sequence model is available in the latest version of Moses

References

Nadir Durrani, Philipp Koehn, Helmut Schmid, Alexander Fraser (2014). Investigating the Usefulness of Generalized Word Representations in SMT. In Proceedings of the 25th Annual Conference on Computational Linguistics (COLING). Dublin, Ireland. August

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, Philipp Koehn (2013). Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In Proceedings of the 51st Annual Conference of the Association for Computational Linguistics (ACL). Sofia, Bulgaria, August.

Nadir Durrani, Alexander Fraser, Helmut Schmid (2013). Model With Minimal Translation Units, But Decode With Phrases. In Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Atlanta, Georgia, USA, June.

Nadir Durrani, Helmut Schmid, Alexander Fraser, (2011). A Joint Sequence Translation Model with Integrated Reordering. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL HLT), Portland, Oregon, USA, June.