# An Exploration of Segmentation Strategies in Stream Decoding

Andrew Finch  Xiaolin Wang  Eiichiro Sumita

Multilingual Translation Group
National Institute of Information and Communications Technology
Kyoto, Japan

# Overview

- Motivation
- Explanation of stream decoding
- Alternative strategies
  - Degree of commitment
  - Minimizing forced monoticity
  - Introducing segmentation cues into the stream
- Conclusion and future directions

# Motivation

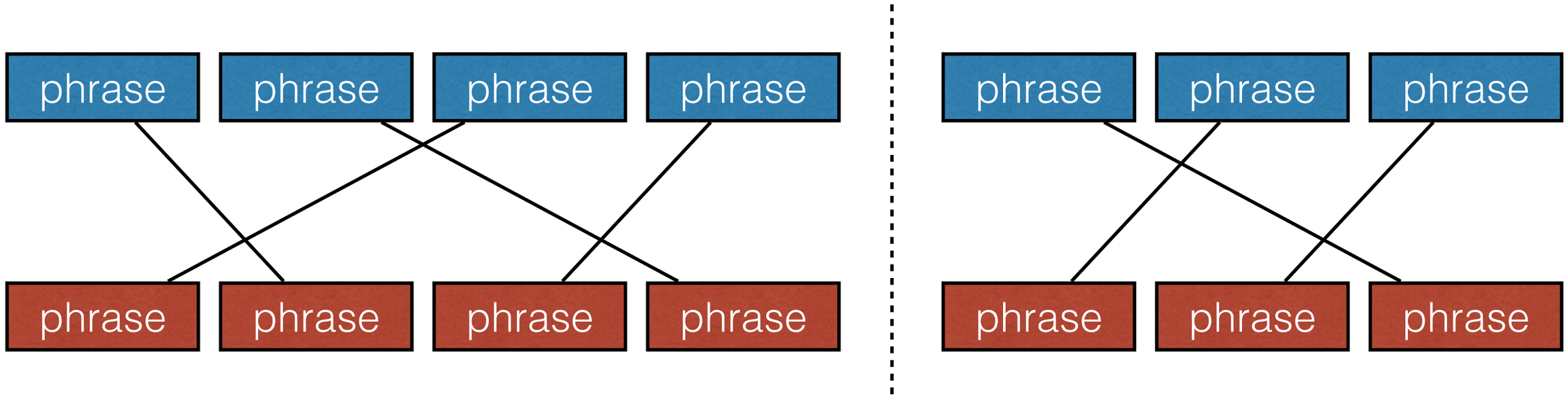
- In interpretation, the interpreter cannot fall too far behind the speaker
- The interpreter must therefore periodically commit to producing a partial translation
- Getting this segmentation right is pivotal in the success of a machine interpretation system
  - It is undesirable, perhaps impossible to take back a partial translation
  - Continuing from a bad position may inevitably lead to a bad translation

# Segmentation Approaches

- Many previous approaches pre-segment the input
  - Fixed length segments
  - Punctuation (Sridhar et al. 2013)
  - Optimizing BLEU (Oda et al. 2014)
  - many others

- Stream decoding (Kolss et al. 2008) segments during the decoding
  - There are segmentation cues in the decoding process

# The Basic Principle




- A possible segmentation point exists if
  - The source phrase before it is part of a contiguous sequence of translated source phrases that extend back to the end of the previous output
- Investigated as a pre-segmentation strategy in Sridhar et al. 2013, but was ineffective

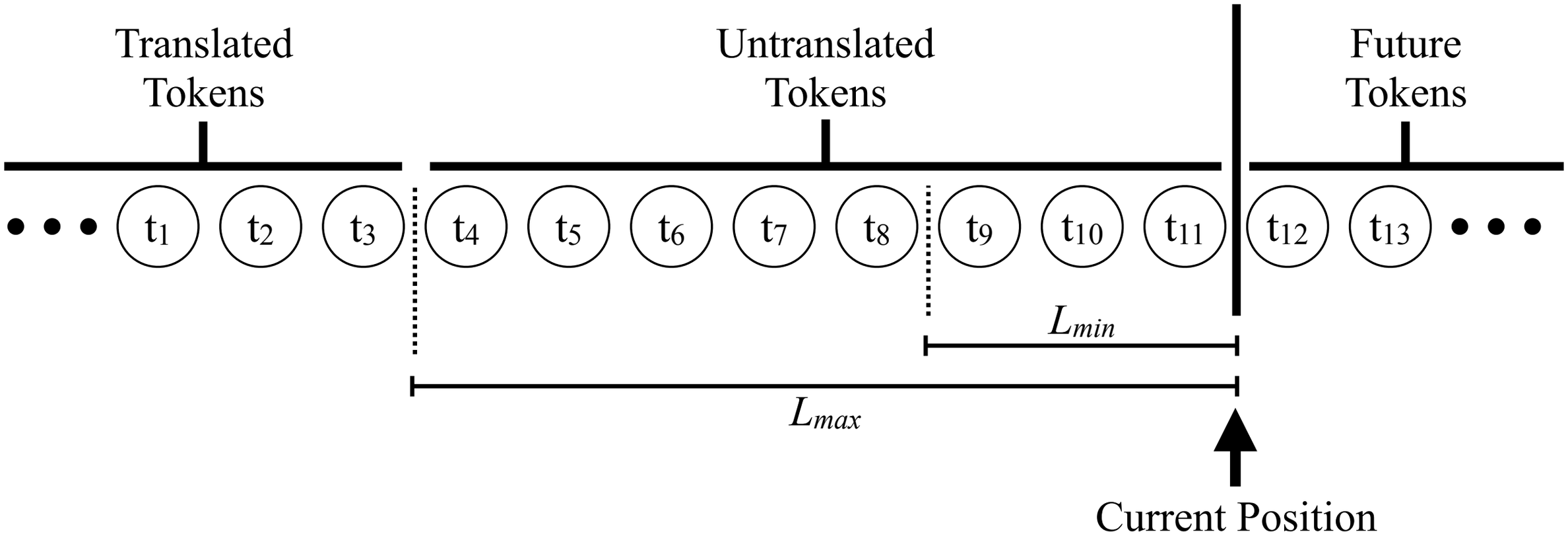
# Stream Decoding Overview

- Phrase-based SMT processes sentence by sentence
- A stream decoder operates on a theoretically infinite stream of tokens (or tuples from a confusion network)
  - A search graph is maintained.
  - As new tokens arrive on the stream, the search graph is extended
  - When the decoder falls too far behind, it is forced to make an output
    - The search graph is truncated, by removing paths irrelevant to the output

# Stream Decoding Parameters

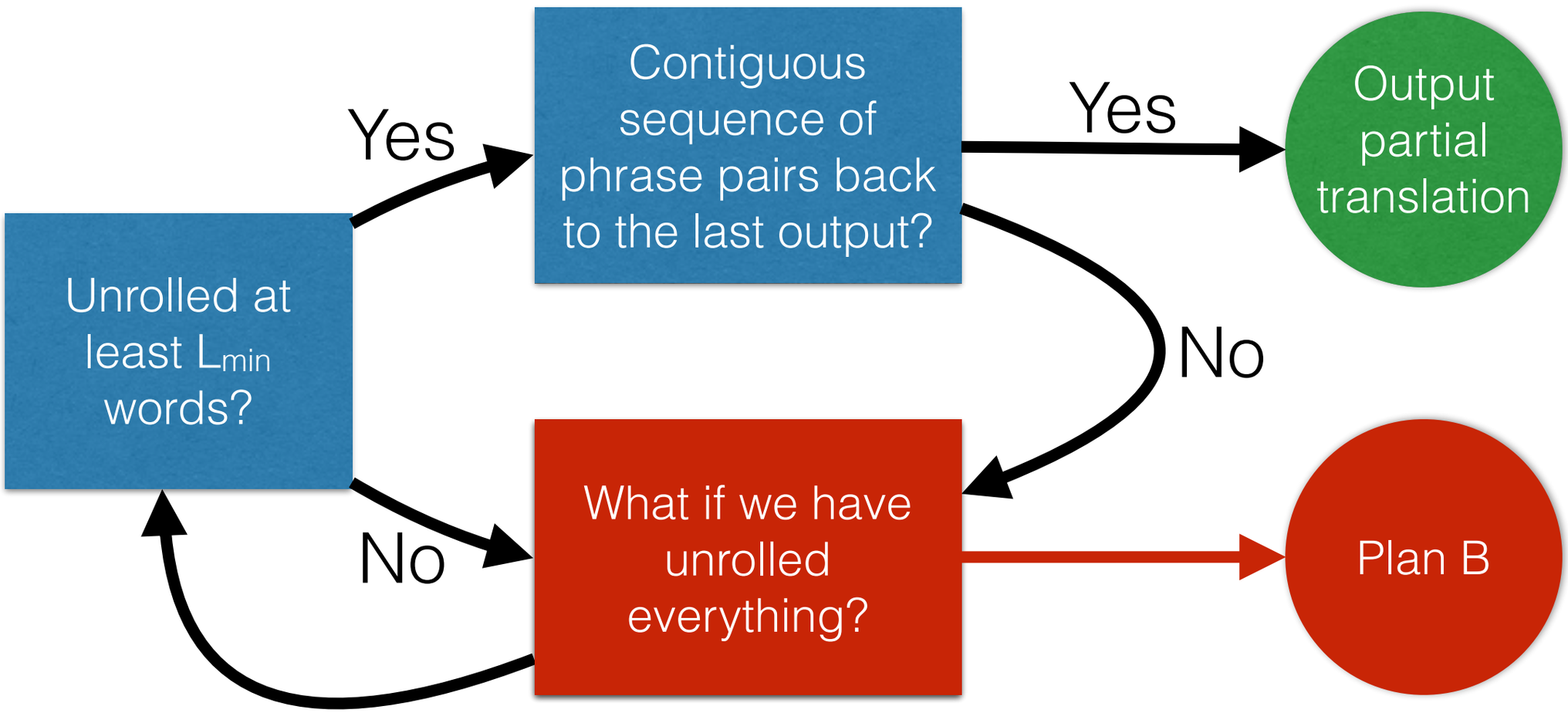
- **$L_{max}$** the number of words the decoder is permitted to fall behind.
  - When the number of untranslated words hits this threshold, the decoder is forced to output something
- **$L_{min}$** the the number of words the decoder should leave untranslated.
  - When the decoder commits to a partial translation, it will not translate the last $L_{min}$ words from the stream.
  - Prevents the decoder from committing too early

# Stream Decoding in Action



# Choosing where to segment: Plan A

- Unroll the best translation hypothesis backwards phrase-by-phrase



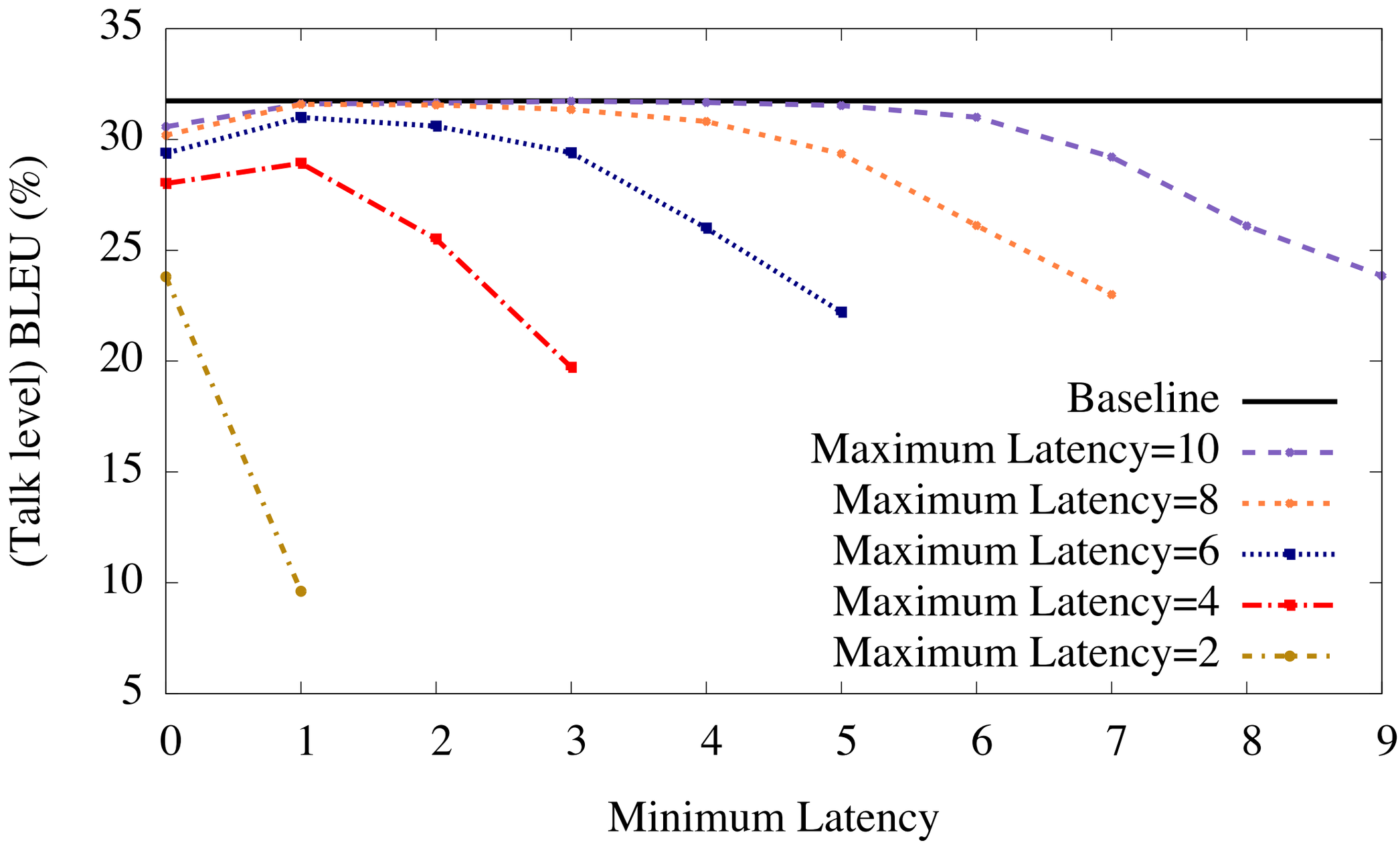
# Plan B

- Re-decode, but this time force the first step to be monotonic..
- .. then execute Plan A.
  - This time success is guaranteed.

# Methodology

- Used the TED talk data from the IWSLT campaign.
- We ran experiments on all the language pairs, but obtained similar results on pairs of European languages.
  - Results on English-Spanish, English-German, and English-Chinese are reported here
- 170-180K ‘sentence’ training set, 887 development set, 1400-1700 test set.
- OCTAVIAN in-house PBSMT decoder
- Evaluated with latest multi-bleu script at the talk-level

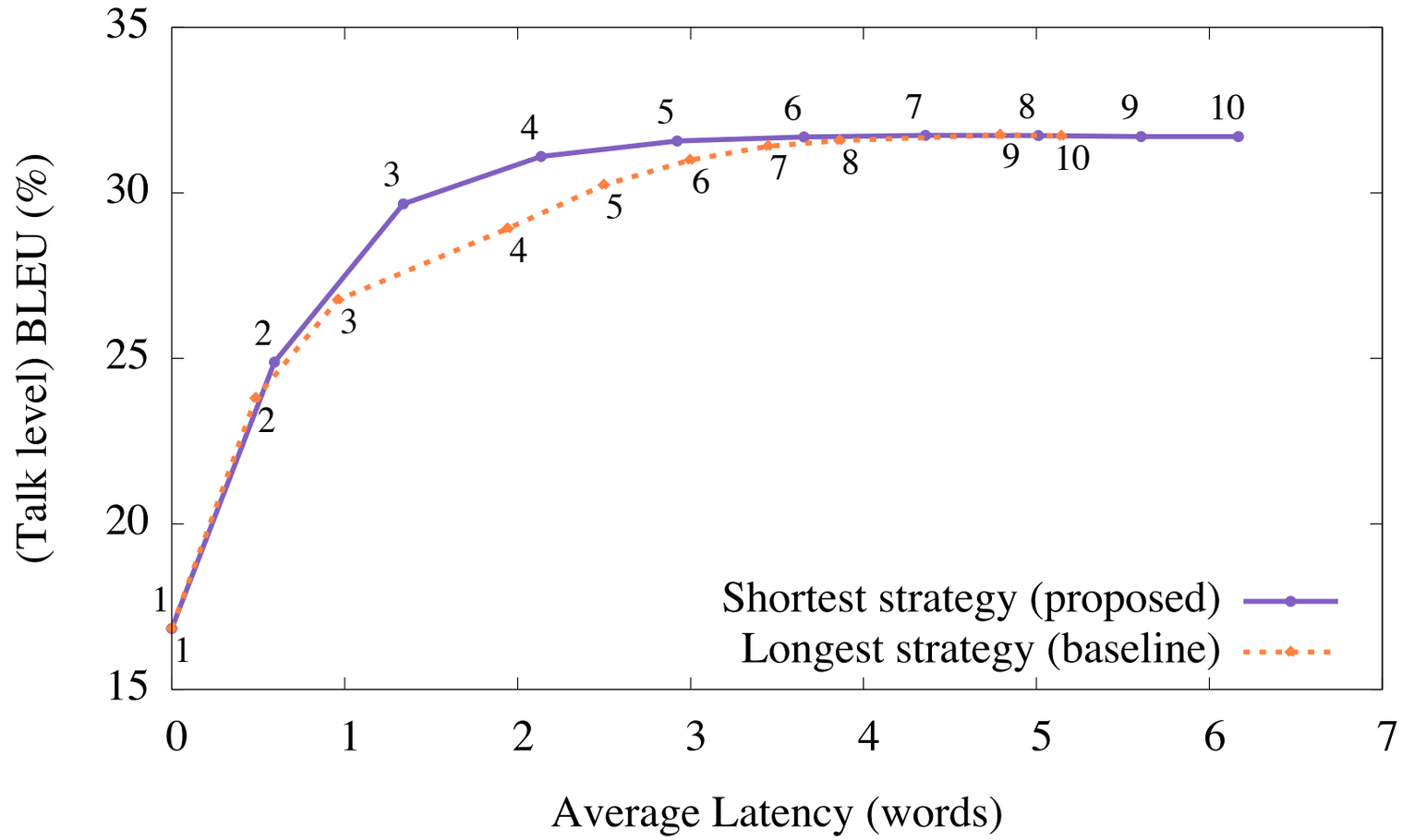
# English-Spanish



# Increasing the output frequency

- The original stream decoder outputs the longest possible translation subject to its constraints.
- Other strategies are possible, for example, it is possible to output the shortest possible translation.
  - Roll back the best hypothesis as far as possible
- Measure the performance with “average latency”: the average number of words each word output is behind the speaker.

# Longest vs. Shortest

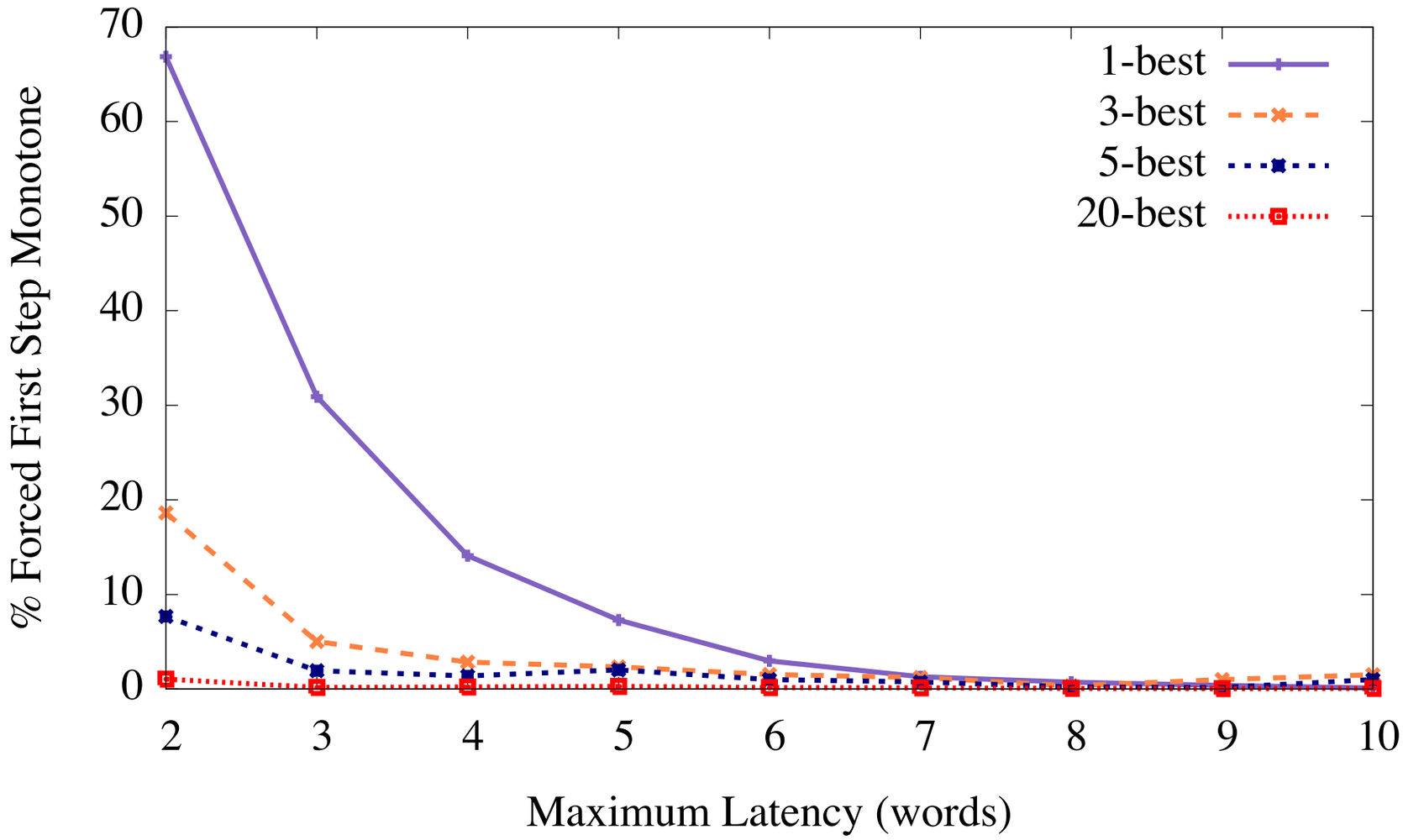



- We observed a small improvement for most language pairs, but for some (for example English-French) there was almost no difference

# Reducing the number of forced monotonic steps

- Forcing the decoder to make a monotonic first step is best avoided and may lead to an unnatural translation.
- Perhaps there are alternative hypotheses in the search graph that can be used to avoid this issue.
- One way is to search the n-best list for a suitable hypothesis.
  - This has the disadvantage that you will proceed with a lower scoring (worse) hypothesis

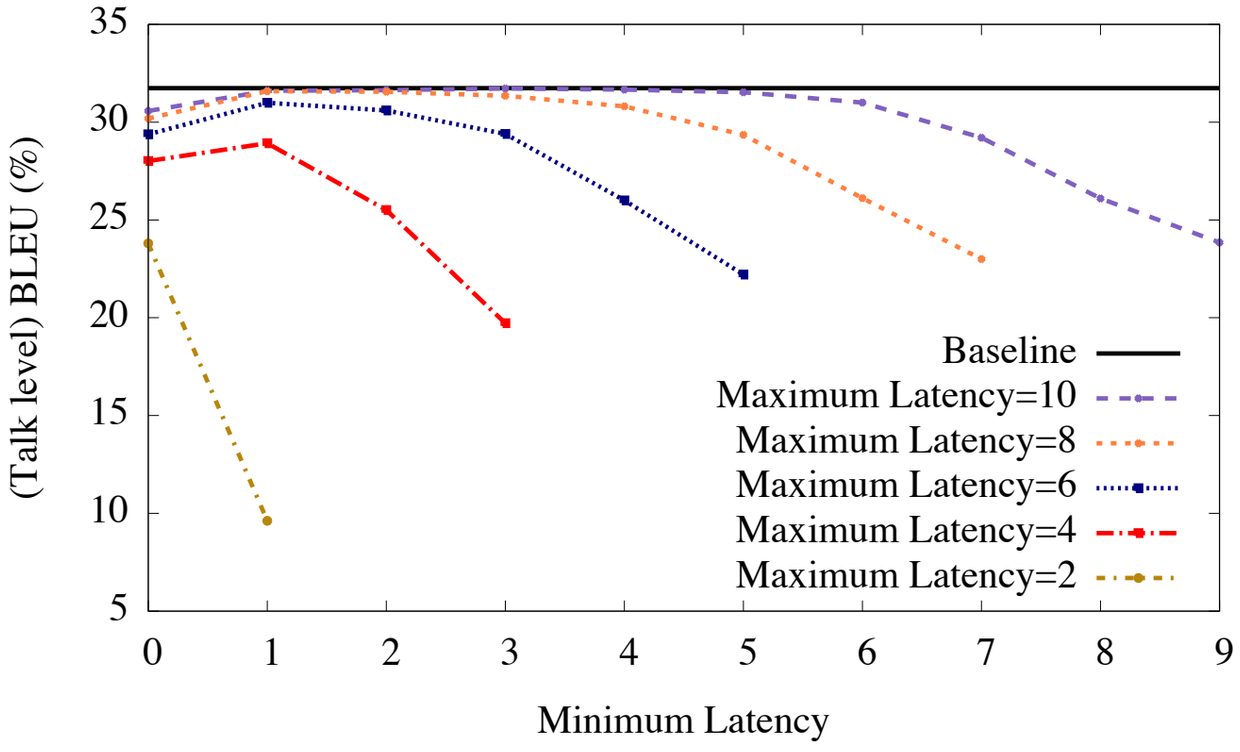
# Using n-best lists



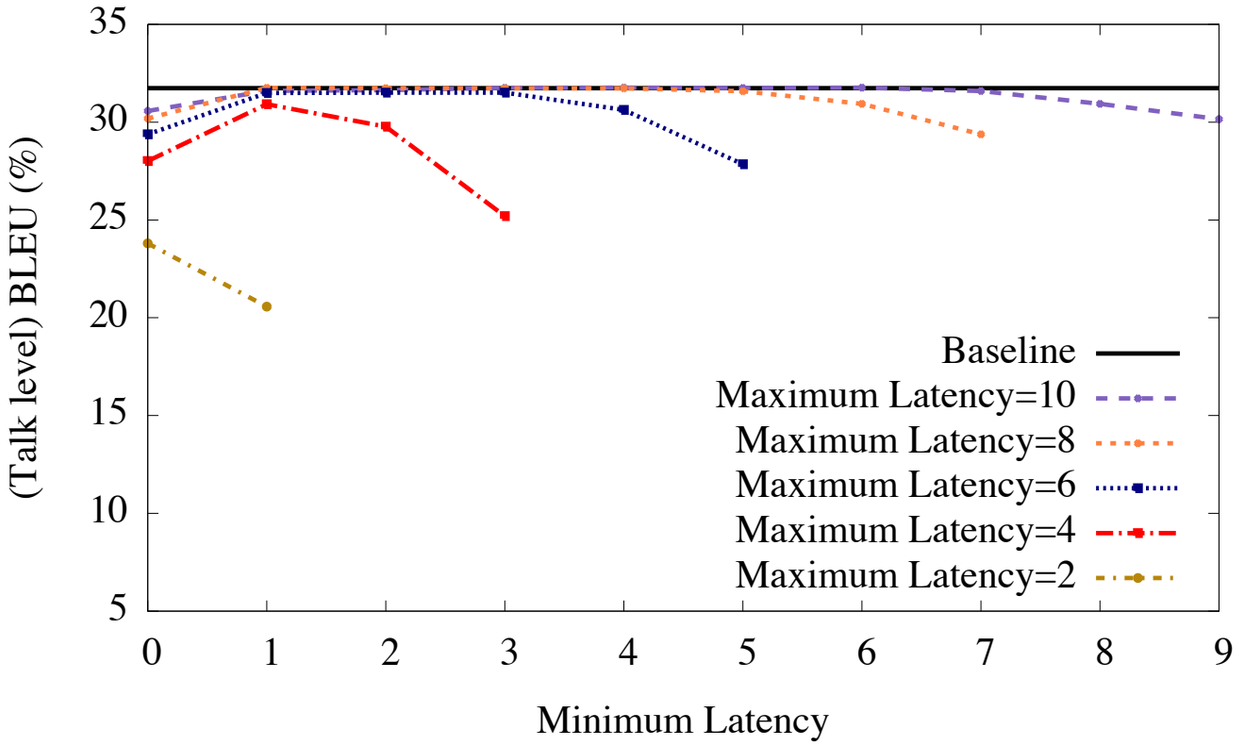

- Searching the n-best list can have a large impact on the number of instances where a forced monotonic is required, especially at lower latencies.

# English-Spanish: 1-best vs. 20-best

1-best

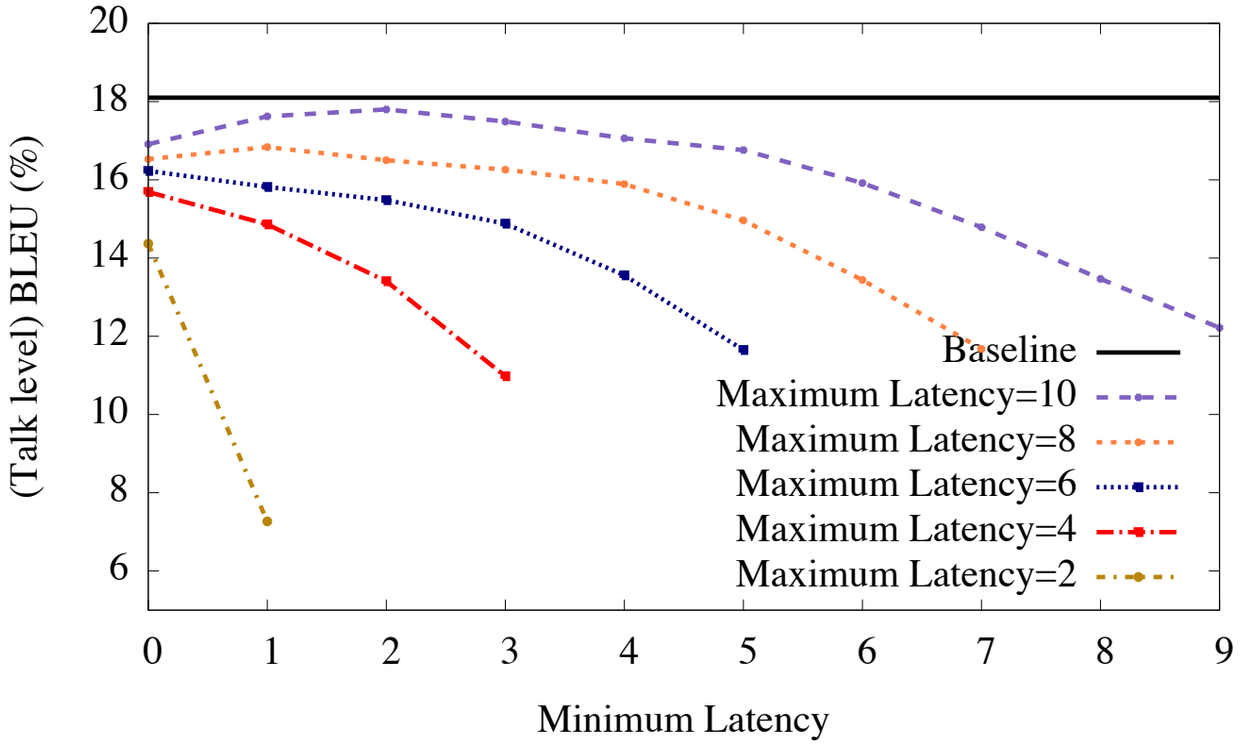



20-best

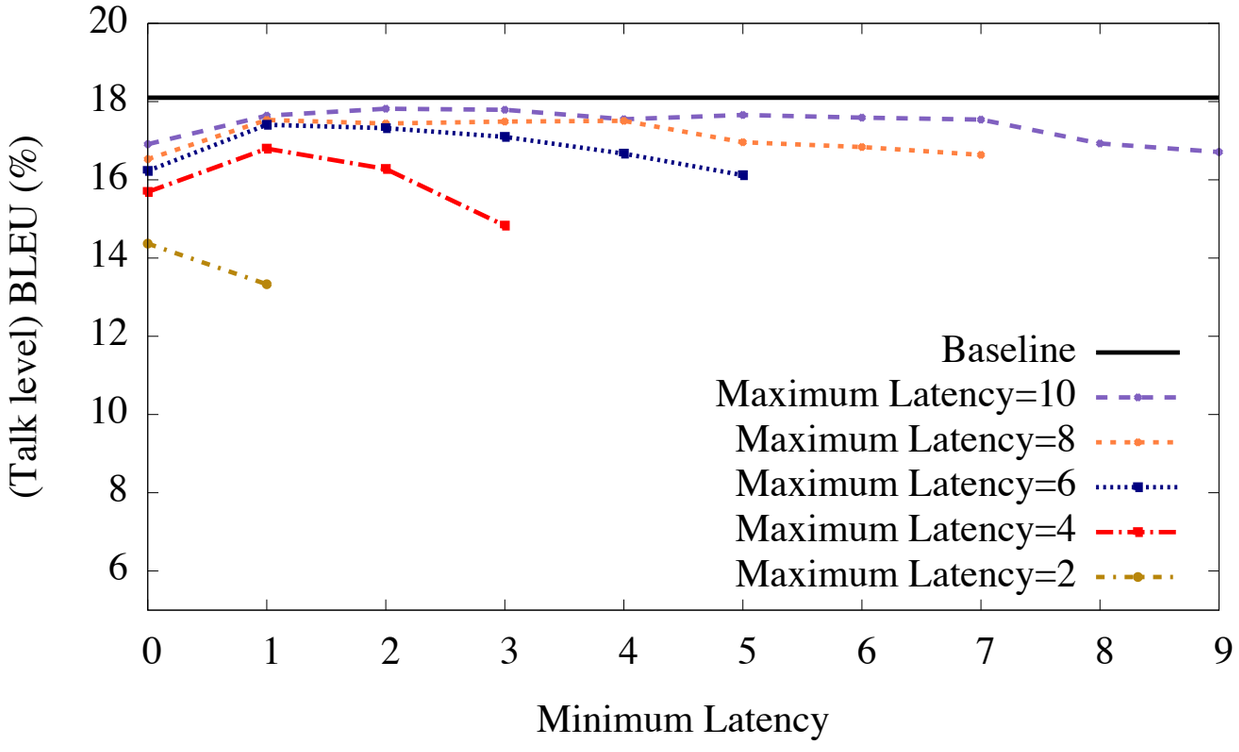


# English-Chinese: 1-best vs. 20-best

1-best



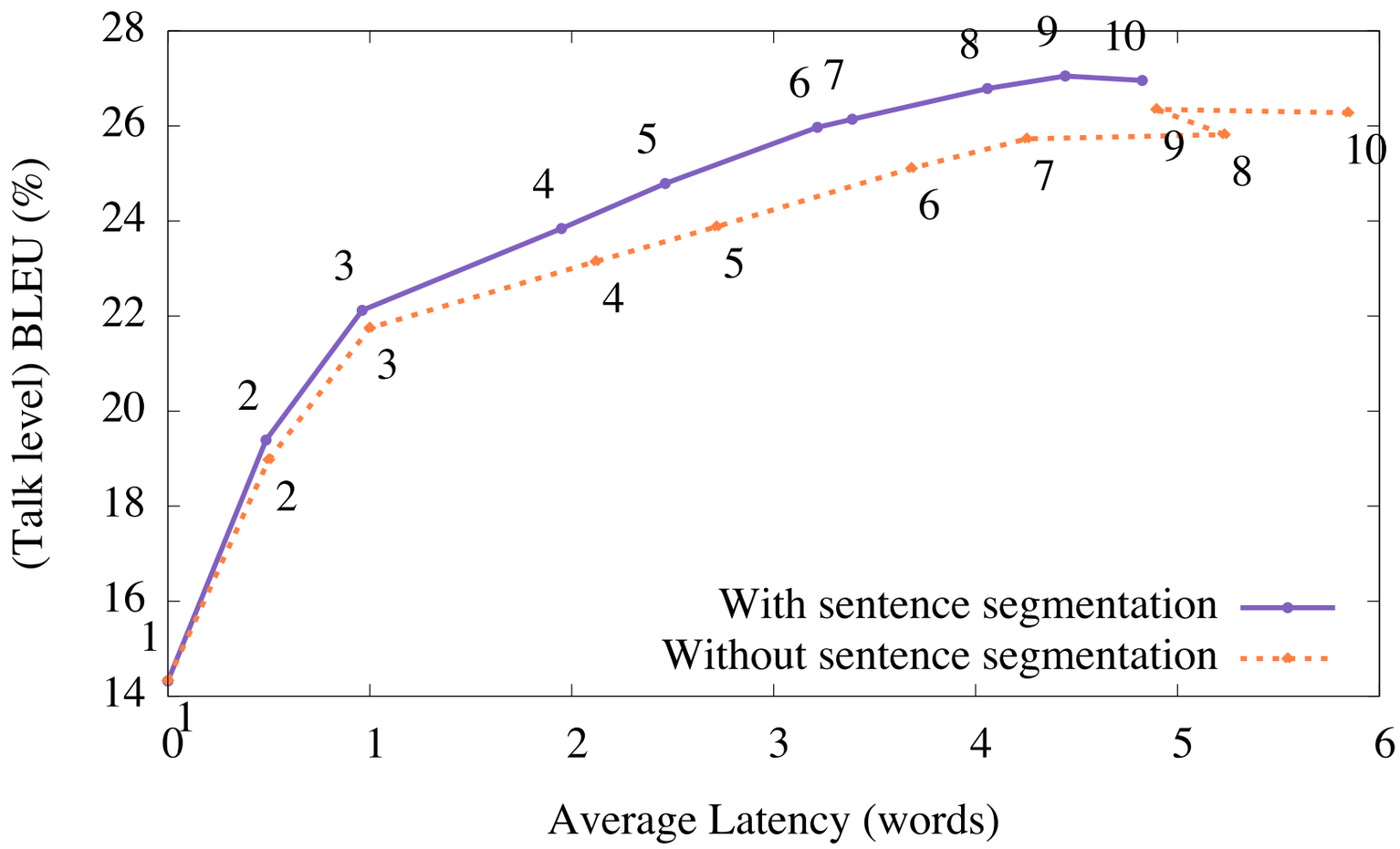

20-best



# Introducing segmentation cues on the stream

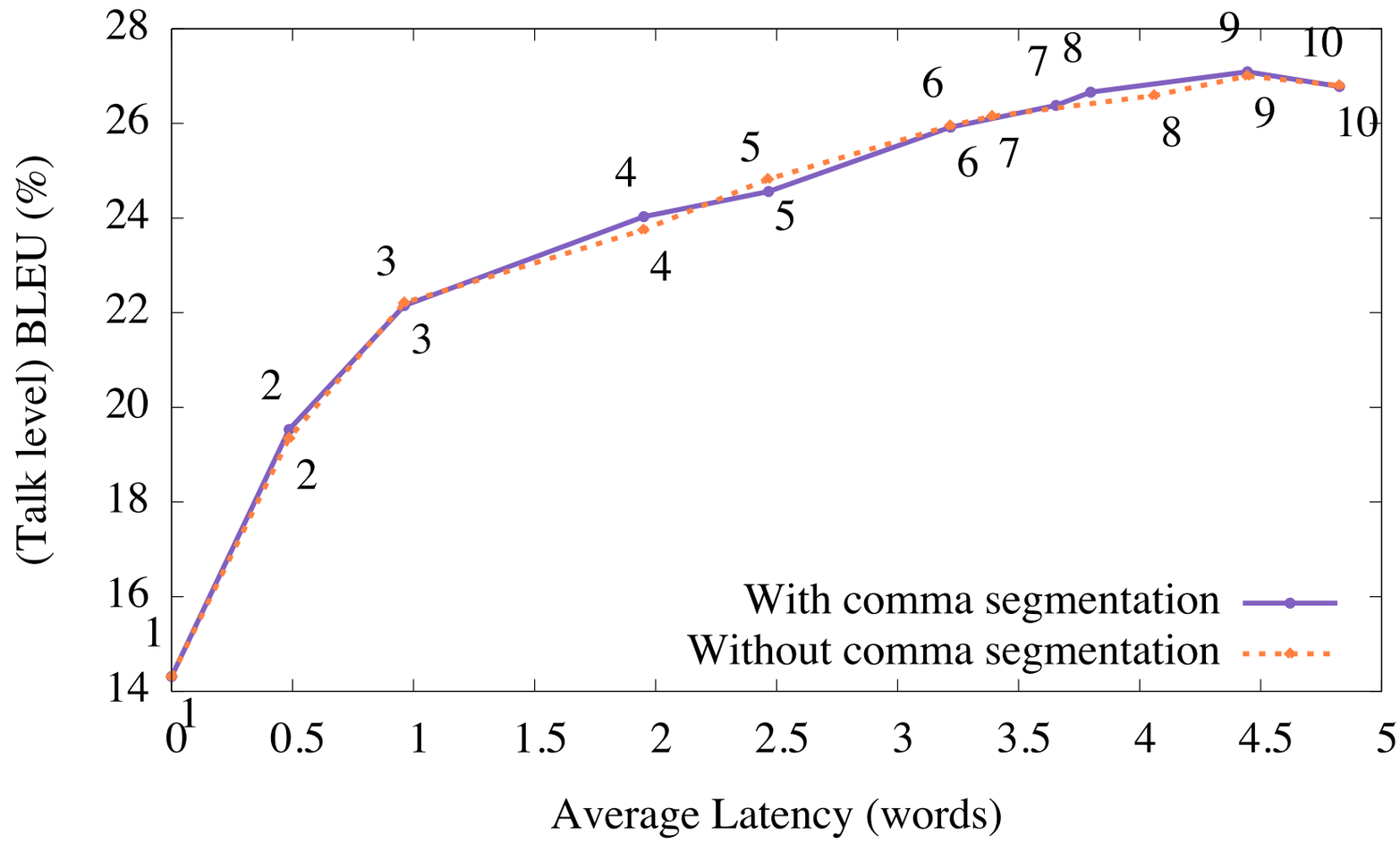
- It is possible to combine stream decoding with approaches that predict the segmentation through other means
- We propose to introduce 2 types of token onto the stream:
  - **<s>** forces the decoder to translate and output up to the token; the language model context is discarded
  - **<p>** forces the decoder to translate and output up to the token; the language model context is preserved

<S>




- The sentence boundary information from the TED corpus was used to annotate the segmentation cues

# <p> given <s>




- Commas in the source TED corpus was used to annotate the segmentation cues
- Commas appear to be ineffective

# Conclusions

- We evaluated the stream decoder on new language pairs and data sets
  - Our findings support the original work
  - Stream decoding can yield high quality translations with low latencies
  - The approach works for languages pairs like English-Chinese with a moderate amount of re-ordering

# Conclusions

- We proposed and investigated alternative decoding strategies:
  - For some pairs it may be better to use a strategy that outputs more frequently
    - It is better to avoid making forced monotonic steps. We proposed a straightforward n-best list extension to the original approach.
- Sentence boundary information can be used effectively by the stream decoder
- Punctuation cannot provide any further benefit

# Future Directions

- Study the application of the steam decoder to language pairs with more serious re-ordering issues, for example English-Japanese
- Study other methods of choosing a single state from the search graph (we tried several variants, but found our simple n-best approach to be the most effective).
- Introduce an statistical segmentation model into the decoder

Thank you for listening!

# English-German