Baseline System Description

Overall Task
- German-to-English translation
- Constrained data condition

Translation model
- Hierarchical phrase-based translation model from provided Europarl and News Commentary parallel training data
- Based on synchronous context-free grammar rules
- Non-terminal span limit was 12 for non-glue grammars
- Grammar was extracted using a suffix array rule extractor
- Segmentation variants of the raw input are encoded in the word lattice
- The decoder automatically chooses which segmentation is best for translation
- In place of translating a single representation of the input, we encode alternative strategies and inference algorithms
- System parameters optimized using VEST, an implementation of minimum error rate training (MERT) within a semiring, on the dev set
- Compute MERT line search using: 

Viterbi Envelope Semiring Training (VEST)

Motivation
- Semiring operations can be overloaded to compute standard quantities in a generic framework
- Useful mathematical abstraction defining two general operations, addition and multiplication

Implementation
- System parameters optimized using VEST, an implementation of minimum error rate training (MERT) within a semiring, on the dev set
- Compute MERT line search using:

Minimum Bayes risk decoding

Motivation
- VEST uses maximum derivation decision rule
- Empirically useful to rescoring with minimum risk decision rule to minimize expected loss

Implementation
- Estimate posterior distribution using a unique 500-best list of hypotheses
- Posterior scaling factor \( \alpha \) was optimized on a held out development set for VEST

Compound Segmentation Lattices

Motivation
- German possesses a rich inflectional morphology, productive compounding, and significant word reordering
- In place of translating a single representation of the input, we encode alternative ways of segmenting compound words in a word lattice
- The decoder automatically chooses which segmentation is best for translation, leading to markedly improved results
- Segmentation variants of the raw input are encoded in the word lattice

Lattice construction
- Create maximum entropy model to model compound word splitting
- Features: frequency of hypothesized morphemes as separate units, number of predicted morphemes, number of letters in a predicted morpheme
- Learn parameters to maximize conditional log-likelihood using a small amount of manually created reference lattices
- Create dev/test lattice of segmentations for words \( \geq 6 \) letters and prune unlikely paths using max-marginals

Bloom Filter LM

Motivation
- LM complexity causes tradeoff between translation quality and decoder memory usage and speed
- Delays caused when LM size necessitates remote language model server
- Randomized language models (RandLM) [4] using Bloom filters can be used locally

Implementation
- Convert existing SRILM directly into a RandLM using default settings

Grammar Extraction

Motivation
- SCFG's are expressive at the cost of a large number (millions) of rules
- Memory requirements have either been for the grammar, when extracted beforehand, or the corpus, for suffix arrays

Implementation
- Sentence-specific grammars extracted and loaded on an as-needed basis by the decoder

References