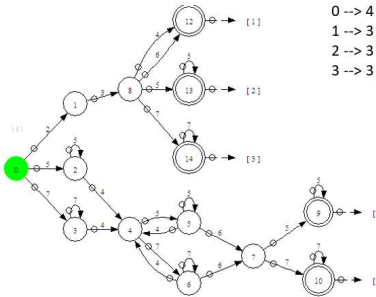
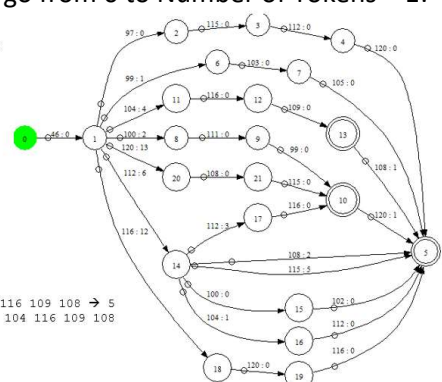
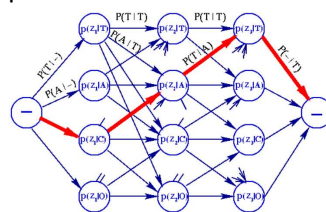


Overview of Tokenization Algorithms

Pattern Based	Word Piece	Unigram LM	BPE
<p>1. Someone manually creates patterns for tokens</p>	<p>1. Trained vocab.txt with "pieces" and patterns for tokens are combined. No scores.</p>	<p>1. vocab.txt is learned by training a Unigram Language Model with expectation maximization rule, scores are probabilities.</p>	<p>1. Starting from characters, vocab.txt is learned by iterative merge of most frequent pair of tokens into one token, scores are frequency rank.</p>
<p>2. Moore Machine is used to find token boundaries and piece boundaries for Word-Piece:</p> <p>1. Char → Eq class: 2 --> 2 4 --> 3 42..47 --> 4 48..57 --> 5 61 --> 6 65..122 --> 7</p> <p>2. Min Moore DFA:</p>  <p>3. Action map: 0 --> 4 0 0 0 4 1 --> 3 0 0 1 2 --> 3 0 0 2 3 --> 3 0 0 3</p> <p>Action map contains IDs and optional context information</p>		<p>2. Mealy Machine maps a Token into ID and ID into Token, such that ids go from 0 to Number of Tokens – 1:</p> <p>1. Char → Eq class: 2 --> 2 4 --> 3 42..47 --> 4 48..57 --> 5 61 --> 6 65..122 --> 7</p>  <p>3. Action map: 0 --> 2 -0.012 2 1 --> 2 -0.022 10 2 --> 2 -0.312 21 3 --> 2 -0.002 3 ...</p> <p>46 104 116 109 108 → 5 5 → 46 104 116 109 108</p> <p>Action map contains Scores and IDs</p>	
<p>3. Find all matches in the string using the Deterministic Finite State Machine (we also implemented several fast forward heuristics to speed up the matching process)</p>			
<p>4. N/A</p>	<p>4. Check that a token is covered completely by pieces otherwise output UNK id for this token.</p>	<p>4. If a character is unknown it creates an unknown token, if before this unknown token there is another unknown token then they are merged into one unknown token.</p>	
<p>5. N/A</p>	<p>5. One globally optimal from start to end sequence is found in the graph of all possible breaks:</p> 		<p>5. A greedy approach is applied. Sort all matches by the frequency rank and apply one after another until no more can be applied.</p>
<p>6. Result: One sequence</p>			