|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Paper | Language + Corpora | What is analysed | How | Features | Baseline |
| Fitzgerald, Hall, Jelinek 2009 | English, SSR | sentences where reparandum can simply be deleted: filler, rough copy, non-copy | Conditional Random Fields (CRFs): GRMM package (Sutton, 2006) | lexical: lexical item, POS, distance to next matching POS, partial word, position in sentence  JC04: is word edit/filler?  LM: unigram & n-gram probabilities | JC04 model (Johnson and Charniak, 2004) |
| Ferguson, Durrett, Klein 2015 | English, Switchboard | Deleting disfluent chunks | Semi-Markov conditional random field (semi-CRF) | token: repeated words or POS, distance to repeated word or POS, duplicate bigram  span: length of span + is proposed span repeated, number of uncommon words contained in the span  acoustic: pauses, word duration (deviation from average)  ToBI: break indices | CRF model |
| Shriberg 2000 | English, Broadcast News + Switchboard | automatic speech segmentation | Decision tree model | prosodic: pause durations, phone durations, pitch information, and voice quality information | word-based statistical models |
| Liu, Shriberg 2003 | English, Switchboard-I | detection of interruption points (Ips) and start of disfluencies  No word fragments used | Decision tree  combination of prosody and word- and POS based LM models: interpolate the posterior probabilities from all three models: identify IPs  Rule-based knowledge: identify start of disfluency | prosodic: word dur, pause dur, phone dur, F0  voice quality: jitter, spectral tilt, open quotient  LMs: hidden-event word + POS based, repetition pattern |  |
| Dufour 2009 | French Broadcast News | characterisation of spontaneity level: prepared speech, low spontaneity or high spontaneity |  | vowel duration, lengthening of word-final syllable, speech rate, number of filler words, number of 1-gram and 2-gram repetitions, bags of n-grams on words, POS tags and syntactic chunk categories |  |
| Audhkhasi 2009 | English, custom corpus, shared on request | filled pauses | SD of formants over a window | formants stay stable and do not exhibit effects of coarticulation: first two formants | cepstral variation (MFCCs) + pitch based | |
| candea-vasilescu-adda\_2005 |  | Do fillers have similar features across different languages? |  |  |  |
| Hugh, Purver 2013 | Switchboard | Incremental disfluency detection with RNNs | RNN | No feature engineering, only 2- or 3- word windows with/without POS tag |  |
| Qian, Liu 2013 | Switchboard |  | Max-Margin Markov Networks (M³Ns) |  |  |
| Johnson, Charniak 2004 | Switchboard | Identify and correct repairs in speech transcripts | Noisy channel model: syntactic parser = source model, TAG-based transducer = channel model |  |  |
| Wang, Che, Liu 2016 | Switchboard + Chinese corpus | Detecting repair type disfluencies | BI-LSTM |  | CRF |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Paper | Language | What is analysed | How | Result |
| Schober 2009 | English | How do listeners detect and process disfluencies | Listeners were asked to start an event based on prompts: fluent speech, (mid-word) interruption with (filled) pause | - Listeners responded equally quickly to the target (repair) word whether the preceding reparandum was interrupted mid- or between-word  - with a longer editing interval (allowing more time to process the evidence that there is some trouble), listeners are better able to process the repair and select the correct target word more quickly |
| Liu, Shriberg 2005 |  | Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection |  |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Corpus name | Paper | Language | Size | Annotation |
| KIT | Cho, Fünfer, Stüker, Waibel 2014 | German | 104K token: 6% filler, 3% rough copy (375 partial words), 1.3% non-copy | Filler words, repetitions and corrections (rough copy), false starts and aborted sentence (non-copy) |
| SSR (Spontaneous Speech Reconstruction) | Fitzgerald, Jelinek 2008 | English | 6.4 K sentences | reconstructed sentences aligned with original: delete, insert, substitute, move, add/remove sentence boundary |
| French Broadcast News |  | French | 11h |  | |
|  |  |  |  |  |
|  |  |  |  |  |

Kaldi e2e training:

<https://github.com/kaldi-asr/kaldi/blob/master/egs/iam/v2/local/chain/tuning/run_e2e_cnn_1a.sh>

<https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/local/chain/e2e/run_tdnn_flatstart.sh>

This is probably what should be used:

<https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/local/chain/e2e/run_tdnn_lstm_flatstart.sh>

script: wsj/s5/local/chain/e2e/ run\_tdnn\_lstm\_flatstart.sh

called from run\_e2e\_char.sh