A Readability Metrics

The readability features used were:

Average sentence length Part of the grade level feature

Average syllable count Part of the grade level feature

Average synset count Potential meanings of word

Link in co-reference graph References to same entity

Cross entropy Typicalness of the passage

Depth (dependency graph) Number of steps between two related words

Maximum dependency length Distance between related words

Depth (parse tree) Estimate of complexity of sentence

OOV (CMU) Word missing from CMU dict when looking for syllables

OOV (WN) Word missing from WordNet when looking for synset

Sentence count Length of passage

VP count Number of clause in text

Table 8 shows their performances, as well as the performance of a logistic regression model combining them.

B Density, Proximity, and Acceptability Metrics

The density proximity and acceptability metrics used were:

Density (aka Context) After matching words in the passage and word from the question, count the number of matches within C words of the ith word in the passage. That is, $C_i = \sum_{j=i-C}^{i+C} m_j$, where m_j is 1 if the jth passage word is a Q-word, and 0 otherwise.

Entropy Proxy for the diversity of match counts in a context. If all words have similar match counts, then the match counts' distribution would have a spike, resulting in a low entropy. If the match counts are distributed more uniformly, the entropy would be higher.

Proximity Get the distance between an answer and the nearest maximum value of C_i. If the answer contains a peak, this value is zero.

Nearest Get the distance between an answer and the nearest match. If the answer contains a match, this value is zero.

Spread Get the number of occurrences of the maximum value of C_i.

Typed Binary feature indicating if at least one human answer is a named entity.

Equiv Number of potential answers with the same averaged C_i. If Typed is true, other named entity of the same type are potential answers. Otherwise, all words are potential answers.

Comp Number of potential answers.

Rank Rank of the human answer among all the potential answers.

Position Index of the first word in the human answer. Used to potentially break ties, as the systems seemed biased towards the first potential answer.

Table 9 shows their performance, as well as the performance of a logistic regression model combining them.

Metric	Easy	Hard	Hard vs Easy
Combined	0.54	0.52	0.54
Average sentence length	0.50	0.52	0.52
Average syllable count	0.52	0.52	0.53
Average synset count	0.52	0.50	0.51
Link in co-reference graph	0.51	0.54	0.54
Cross entropy	0.51	0.54	0.54
Depth (dependency graph)	0.51	0.52	0.53
Maximum dependency length	0.50	0.51	0.51
Depth (parse tree)	0.52	0.52	0.52
OOV (CMU)	0.51	0.50	0.50
OOV (WN)	0.51	0.52	0.52
Sentence count	0.50	0.54	0.54
VP count	0.52	0.51	0.52

Table 8: Area under the curve of readability metrics, when detecting easy and hard questions, and when discriminating between easy and hard question. Higher is better, 1.0 corresponds to perfect classification, and 0.5 to random choice.

Metric	Easy	Hard	Hard vs Easy
Combined	0.71	0.67	0.74
competition	0.65	0.54	0.61
comp (NER)	0.59	0.62	0.64
density	0.60	0.66	0.70
proximity	0.59	0.66	0.69
entropy	0.46	0.44	0.43
equiv	0.65	0.64	0.69
equiv (NER)	0.56	0.60	0.61
nearest	0.52	0.51	0.51
pos	0.53	0.51	0.52
rank	0.62	0.65	0.70
rank (NER)	0.61	0.68	0.70
spread	0.55	0.57	0.59
typed	0.63	0.54	0.60

Table 9: Area under the curve of density, proximity and acceptability metrics, when detecting easy and hard questions, and when discriminating between easy and hard question. Higher is better, 1.0 corresponds to perfect classification, and 0.5 to random choice. Some metrics are based on named entities; Performance of the classifier is also measured for the subset of question where the answer is a NE, indicated as “(NER)”.