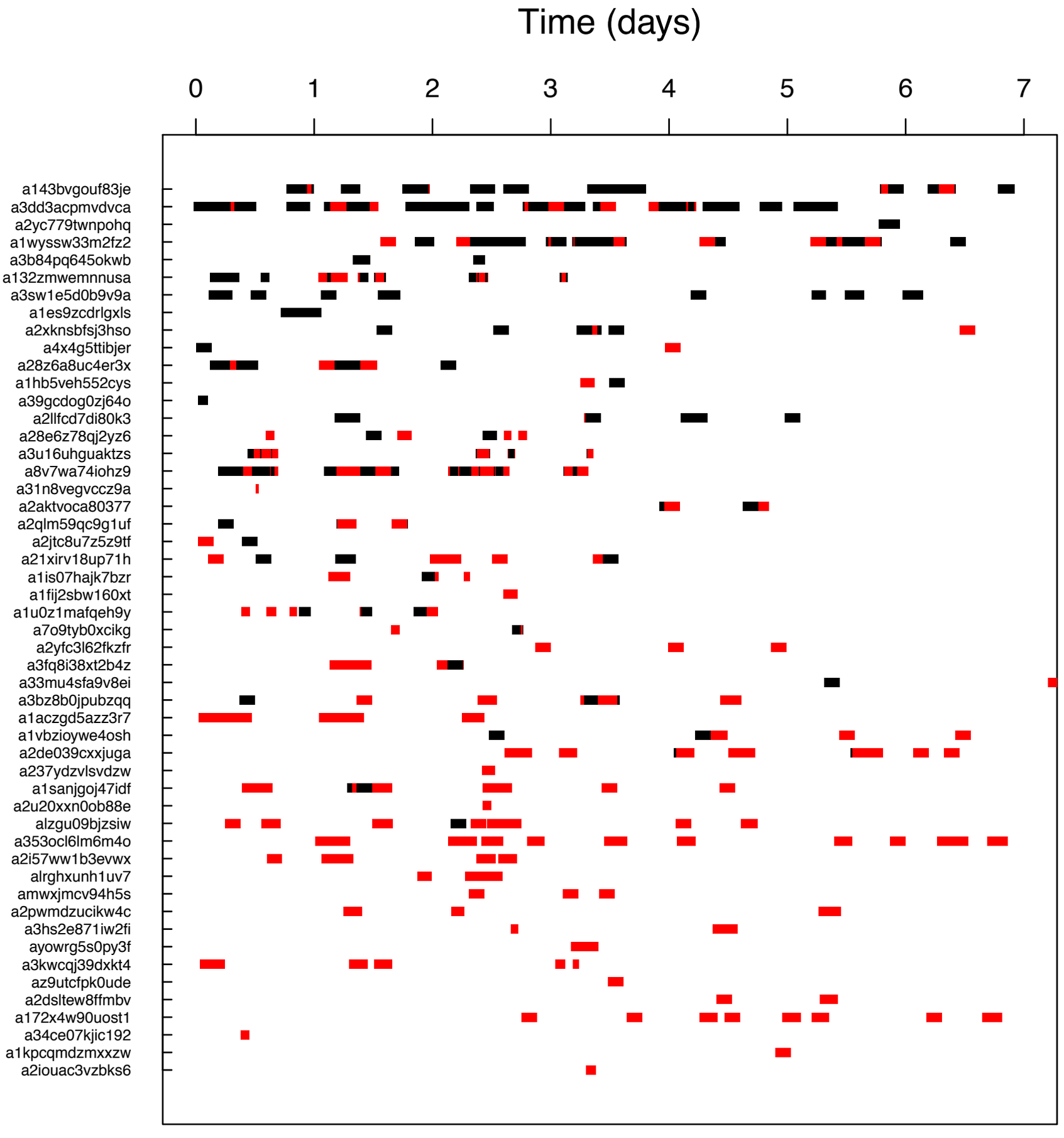

Time (days)
0
1
2
3
4
5
6
7
a143bvgouf83je
a3dd3acpmvdvca
a2yc779twnpohq
a1wyssw33m2fz2
a3b84pq645okwb
a132zmwemnnusa
a3sw1e5d0b9v9a
a1es9zcdrlgxls
a2xknsbfsj3hso
a4x4g5ttibjer
a28z6a8uc4er3x
a1hb5veh552cys
a39gcdog0zj64o
a2llfcd7di80k3
a28e6z78qj2yz6
a3u16uhguaktzs
a8v7wa74iohz9
a31n8vegvccz9a
a2aktvoca80377
a2qlm59qc9g1uf
a2jtc8u7z5z9tf
a21xirv18up71h
a1is07hajk7bzr
a1fij2sbw160xt
a1u0z1mafqeh9y
a7o9tyb0xcikg
a2yfc3l62fkzfr
a3fq8i38xt2b4z
a33mu4sfa9v8ei
a3bz8b0jpubzqq
a1aczgd5azz3r7
a1vbzioywe4osh
a2de039cxxjuga
a237ydzvlsvdzw
a1sanjgoj47idf
a2u20xxn0ob88e
alzgu09bjzsiw
a353ocl6lm6m4o
a2i57ww1b3evwx
alrghxunh1uv7
amwxjmcv94h5s
a2pwmdzucikw4c
a3hs2e871iw2fi
ayowrg5s0py3f
a3kwcqj39dxkt4
az9utcfpk0ude
a2dsltew8ffmbv
a172x4w90uost1
a34ce07kjic192
a1kpcqmdzmxxzw
a2iouac3vzbks6

puted based on the BLEU against professional translations. Each tick represent a single translation and depicts the BLEU score using two colors. The tick is black if its BLEU score is higher than the median and it is red otherwise. Good translators tend to produce consistently good translations and bad translators rarely produce good translations.

5.2 Evaluating Rankings

We use weighted Pearson correlation (Pozzi et al., 2012) to evaluate our ranking of workers against gold standard ranking. Since workers translated different number of sentences, it is more important to rank the workers who translated more sentences correctly. Taking the importance of workers into consideration, we set a weight to each worker using the number of translations he or she submitted when calculating the correlation. Given two lists of worker scores *x* and *y* and the weight vector *w*, the weighted Pearson correlation ρ can be calculated as:

$$\rho(x, y; w) = \frac{cov(x, y; w)}{\sqrt{cov(x, x; w) cov(y, y; w)}} \quad (1)$$

where cov is weighted covariance:

$$cov(x, y; w) = \frac{\sum_i w_i (x_i - m(x; w))(y_i - m(y; w))}{\sum_i w_i} \quad (2)$$

and m is weighted mean:

$$m(x; w) = \frac{\sum_i w_i x_i}{\sum_i w_i} \quad (3)$$

5.3 Automatically Ranking Translators

We introduce two approaches to rank workers using a small portion of the work that they submitted. The strategy is to filter out bad workers, and to select the best translation from translations provided by the remaining workers. We propose two different ranking methods:

Ranking workers using their first k translations We rank the Turkers using their first few translations by comparing their translations against the professional translations of those sentences. Ranking workers on gold standard data would allow us to discard bad workers. This is similar to the idea of a qualification test in MTurk.

Ranking workers using a model In addition to ranking workers by comparing them against a gold standard, we also attempt to automatically predict their ranks with a model. We use the linear regression model to score each translation and rank workers by their model predicted performance. The model predicted performance of the worker w is:

$$performance(w) = \frac{\sum_{t \in T_w} score(t)}{|T_w|} \quad (4)$$

where T_w is the set of translations completed by the worker w and $score(t)$ is the model predicted score for translation t.

5.4 Experiments

After we rank workers, we keep top-ranked workers and select the best translation only from their translations. For both ranking approaches, we vary the number of good workers that we retain.

We report both rankings' correlation with the gold standard ranking. Since the top worker threshold is varied and since we change the value of k in first k sentence ranking, we have a different test set in different settings. Each test set excludes any items which were used to rank the workers, or which did not have any translations from the top workers according to our rankings.

5.4.1 Gold standard and Baseline

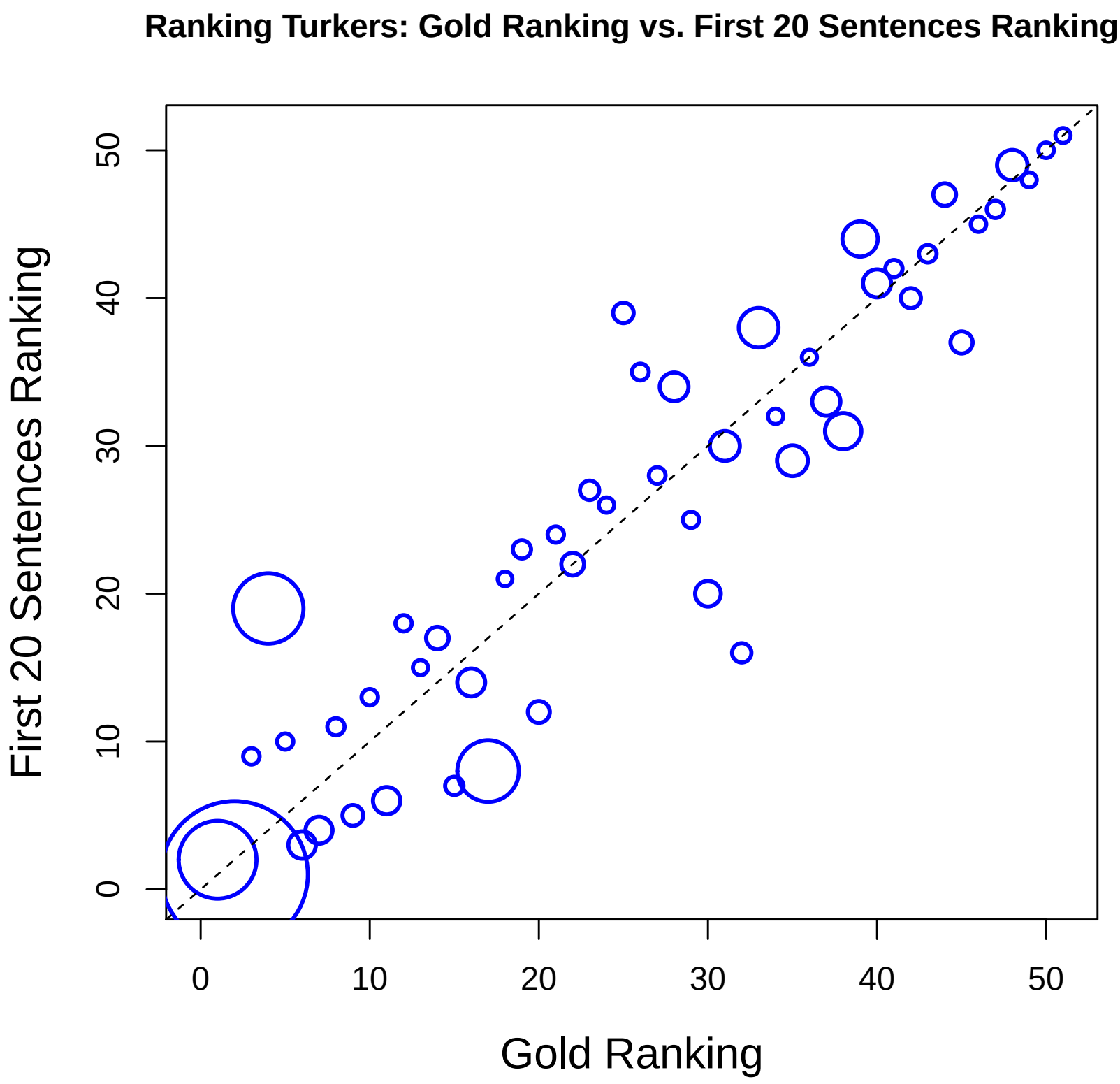
We evaluate ranking quality using the weighted Pearson correlation (ρ) compared with the gold standard ranking of workers. To establish the gold standard ranking, we score each Turker based on the BLEU score comparing all of his or her translations to the corresponding professional references.

We use the ranking by the MERT model developed by Zaidan and Callison-Burch (2011) as baseline. It achieves a correlation of 0.73 against the gold standard ranking.

5.4.2 Ranking workers using their first k translations

Without using any model, we rank workers using their first k translations. We select best translation of each source sentence from the top ranked worker who translated that sentence.

Table 2 shows the results of Pearson correlations for different value of k. As k increases, our rankings


Ranking Turkers: Gold Ranking vs. First 20 Sentences Ranking
First 20 Sentences Ranking
Gold Ranking
0
10
20
30
40
50