Measuring sentence parallelism using Mahalanobis distances: The NRC unsupervised submissions to the WMT18 Parallel Corpus Filtering shared task

Patrick Littell, Samuel Larkin, Darlene Stewart,
Michel Simard, Cyril Goutte, Chi-kiu Lo
National Research Council of Canada
1200 Montreal Road, Ottawa ON, K1A 0R6
`Firstname.Lastname@cnrc-nrc.gc.ca`

Abstract

The WMT18 shared task on parallel corpus filtering (Koehn et al., 2018b) challenged teams to score sentence pairs from a large high-recall, low-precision web-scraped parallel corpus (Koehn et al., 2018a). Participants could use existing sample corpora (e.g. past WMT data) as a supervisory signal to learn what a "clean" corpus looks like. However, in lower-resource situations it often happens that the target corpus of the language is the *only* sample of parallel text in that language. We therefore made several unsupervised entries, setting ourselves an additional constraint that we not utilize the additional clean parallel corpora. One such entry fairly consistently scored in the top ten systems in the 100M-word conditions, and for one task—translating the European Medicines Agency corpus (Tiedemann, 2009)—scored among the best systems even in the 10M-word conditions.

1 Introduction and motivation

The WMT18 shared task on parallel corpus filtering assumes (but does not require) a supervised learning approach. Given

1. a set of "clean" German-English parallel corpora including past WMT data, Europarl (Koehn, 2005), etc., and
2. a large, potentially "dirty" corpus (i.e., one that may contain non-parallel data, non-linguistic data, etc.) scraped from the internet (Koehn et al., 2018a),

can one identify which sentences from (2) are clean? Supervised learning is an obvious approach in well-resourced languages like German and English, in which there exist well-cleaned parallel corpora across various domains.

However, in much lower-resourced languages, we generally do not have *multiple* parallel corpora in a given language pair to assess the quality of the corpus at hand; the corpus to be evaluated is often the only one available.[1] If we want to assess the quality of *one* corpus, we cannot rely on a supervisory signal derived from additional, cleaner corpora. We therefore do not utilize the additional parallel corpora (except as additional sources of monolingual data).

The systems described in this paper were inspired instead by anomaly detection approaches: can we instead attempt to identify sentence pairs that are, in some way, "strange" for this dataset? Considering each sentence pair as a draw from a distribution of high-dimensional vectors, we define an anomalous sentence pair as one whose draw was improbable compared to the probability of drawing its component sentences independently. The resulting measure, conceptually similar to pointwise mutual information albeit couched in terms of Mahalanobis distances rather than actual probabilities, is detailed in §3.

A submission based primarily on this one measurement (with some pre- and post-processing to avoid duplicate and near-duplicate sentences) performed consistently above the median in the 100M-word conditions, and for a few tasks (particularly EMEA translation) was among the top systems even for the 10M-word conditions. It was also the #2 system in one of the dev conditions (WMT newstest2017, NMT trained on 100M words), which is surprising given that it could not have overfit to the development set; it did not utilize the WMT17 development set in any way.

2 Overall architecture

The highest-ranked submission of our unsupervised submissions, `NRC-seve-bicov`,

[1] We are thinking in particular of the English-Inuktitut translation pair, which is a long-standing research interest of NRC (e.g. Martin et al., 2003).

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 913–920
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-6481

shares the same general skeleton as NRC's highest-ranked supervised submission, NRC-yisi-bicov (Lo et al., 2018); it differs primarily in the parallelism estimation component (§2.3).

2.1 Training sentence embeddings

We began by training monolingual sentence embeddings using sent2vec (Pagliardini et al., 2018), on all available monolingual data. This included the monolingual data available in the "clean" parallel training data. That is to say, we did not completely throw out the clean parallel data for this task, we simply used it as two unaligned monolingual corpora.

We trained sentence vectors of 10, 50, 100, 300, and 700 dimensions; our final submissions used the 300-dimensional vectors as a compromise between accuracy (lower-dimensional vectors had lower accuracy during sanity-checking) and efficiency (higher-dimensional vectors ended up exceeding our memory capacity in downstream components).

In a system such as this, which is looking for "strange" sentence pairs, training on additional monolingual data beyond the target corpus carries some risks. If the additional monolingual data were to have very different domain characteristics (say, mostly religious text in the first language and mostly medical text in the second), then the two vector spaces could encode different types of sentence as "normal". On the other hand, not using additional monolingual data carries its own risks; monolingual data that *is* domain-balanced could help to mitigate domain mismatches in the target parallel data (say, newswire text erroneously misaligned to sequences of dates).

2.2 Pre-filtering

Although the input data had already been deduplicated by the shared task organizers, we did an additional de-duplication step in which email addresses and URLs were replaced with a placeholder token and numbers were removed, before deciding which sentences were duplicates. We had noticed that large amounts of data consisted of short sentences that were largely numbers (for example, long lists of dates). Although these sentences were indeed unique, we noticed that several of our parallelism measurements ended up preferring such sentences to such an extent that the resulting MT training sets were disproportionately dates, and performed comparatively poorly when tasked with training full sentences. To mitigate this, we ran an additional de-duplication step on the English side in which two sentences that differ only in numbers (e.g., "14 May 2017" and "19 May 1996") were considered duplicates.

Without numerical de-duplication, we believe the parallelism estimation step in §2.3 would have had too much of a bias towards short numerical sentences. It is, after all, essentially just looking for sentence pairs that it considers likely given the distribution of sentence pairs in the target corpus; if the corpus has a large number of short numerical sentences (and it appears to), the measurement will come to prefer those, whether or not they are useful for the downstream task.

The additional de-duplication also had a practical benefit in that the resulting corpus was much smaller, allowing us to perform calculations in memory (e.g., that in §3.2) on the entire corpus at once rather than having to approximate them in mini-batches.

We also discarded sentence pairs that were exactly the same on each side, in which one sentence contained more than 150 tokens, in which the two sentences' numbers did not match, or in which there were suspiciously non-German or non-English sentences according to the pyCLD2 language detector[2]. When pyCLD2 believed a putatively German sentence to be something other than German with certainty greater than 0.5, or a putatively English sentence to be something other than English with certainty greater than 0.5, it was discarded.

2.3 Parallelism estimation

With sentence vectors (§2.1) for the reduced corpus (§2.2) in hand, we set out to estimate the degree of parallelism of sentence pairs. A novel measure of parallelism, based on ratios of squared Mahalanobis distances, performed better on a synthetic dataset than some more obvious measurements, and the single-feature submission based on it was our best unsupervised submission.

We also made several other unsupervised measurements:

[2] https://github.com/aboSamoor/pycld2

1. Perplexity of the German sentence according to a 6-gram KenLM language model[3] (Heafield, 2011)
2. Perplexity of the English sentence according to a 6-gram KenLM language model
3. The ratio between (1) and (2), to find sentences pairs that contain different amounts of information
4. Cosine distances between German and English sentence vectors, in a bilingual `sent2vec` space trained only on the target corpus

As we did not have a supervisory signal, we did not have a principled way of choosing weights for these features. Instead, we simply took an unweighted average of the above four features and the Mahalanobis feature in §3.2, after rescaling each to the interval [0.0, 1.0]. As seen in §5, systems based on this feature combination (`NRC-mono-bicov` and `NRC-mono`) were outperformed by our single-feature system in most conditions.

We also considered combinations of these unsupervised measurements with supervised measurements, but this attempt was also unsuccessful compared to a system that used only a single supervised measurement for sentence pair ranking (Lo et al., 2018).

2.4 Post-filtering

After scoring each sentence for parallelism, we performed another de-duplication step. In this step, we iterated over each target-language sentence in order of parallelism (that is, sentences assessed to have the highest parallelism were considered first), and removed pairs that only consisted of bigrams that had already been seen. (That is to say, a sentence pair was kept only if it contains a bigram that had not previously been seen.)

This step has to occur *after* quality assessment because, in contrast to regular de-duplication, the sentences in question are not identical; the sentence (and the pair it comes from) may differ in quality from the sentence(s) that make it a duplicate, so we want to keep the *best* such sentence, not just the one that happened to come first in the original corpus.

3 Mahalanobis ratios for parallelism assessment

As mentioned in §2.3, we performed several unsupervised measurements on each sentence pair; of these, the measurement that best predicted parallelism (on synthetic data and on our small 300-sentence annotated set) was a novel measurement based on squared Mahalanobis distances.

This measurement rests on two insights:

- If sentence vectors (or in our case, sentence-pair vectors) are normally distributed, the probability that we draw a particular vector (or a more extreme vector) is related to the squared Mahalanobis distance via the χ^2 distribution.
- If the two sentences relate the same information, the probability of drawing the vector for that pair should not be much less than the probability of drawing the individual sentence vectors in isolation.

While Mahalanobis distance is a common statistical measurement, particularly in anomaly detection (e.g. Reed and Yu, 1990), it is not commonly used in machine translation, so we briefly introduce it below.[4]

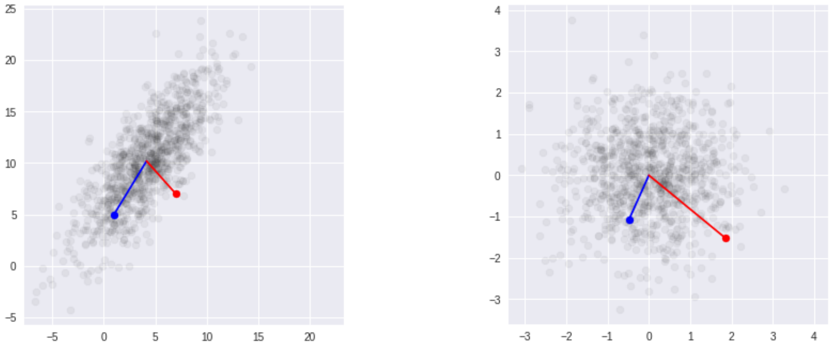
3.1 Mahalanobis distance

The probability of a draw from a univariate normal distribution can be related to its distance to the mean in terms of standard deviations (the z-score). In a multivariate normal distribution, however, just measuring the Euclidean distance to the mean can lead to incorrect conclusions; visual inspection of Figure 1a illustrates that the red vector, despite being a clear outlier, is nonetheless closer to the mean than the blue vector.

Rather, the appropriate measurement for relating distance to probability is the square of the Mahalanobis distance (Mahalanobis, 1936); for a vector x from distribution X with correlation Σ and mean μ:

[3] Although we assumed that high perplexity sentences would be worse—that they might be ungrammatical, for example—sanity checking suggested higher-perplexity sentences were actually better. Error analysis later suggested that many non-parallel (or parallel but non-informative) sentences were short, possibly explaining why taking perplexity as a *positive* feature resulted in higher scores in sanity-checking.

[4] The following relies heavily on the explanation in Boggs (2014). Note that this explanation is also concerned with the square of the Mahalanobis distance rather than the Mahalanobis distance; it is typical for authors to describe both as "Mahalanobis distance" in prose (cf. Warren et al., 2011, p. 10). It is also typical to use "Mahalanobis distance" to specifically refer to Mahalanobis distance from a point to the mean, although this distance is defined for any two points.



(a) Euclidean distance

(b) Mahalanobis distance

Figure 1: Euclidean distance to the mean in a multivariate normal distribution is not necessarily related to probability; in figure (a), the red vector, despite being an outlier, is closer to the mean. In figure (b), we have rescaled and decorrelated the distribution; Euclidean distance measured in the resulting space (the Mahalanobis distance) can be related to probability through the χ^2 distribution.

$$d^2(x) = (x-\mu)^T\Sigma^{-1}(x-\mu) \tag{1}$$

This is equivalent to decorrelating and rescaling to unit variance in all dimensions, via the inverse square root of the correlation matrix ("Mahalanobis whitening"), and then measuring the squared Euclidean distance to the mean in the resulting space.

$$\begin{aligned} d^2(x) &= (x-\mu)^T\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(x-\mu) && (2)\\ &= (\Sigma^{-\frac{1}{2}}(x-\mu))^T(\Sigma^{-\frac{1}{2}}(x-\mu)) && (3)\\ &= \|\Sigma^{-\frac{1}{2}}(x-\mu)\|_2^2 && (4) \end{aligned}$$

Figure 1b illustrates the same distribution transformed by $\Sigma^{-\frac{1}{2}}$; we can see that now the magnitude of the outlier red vector is greater than the magnitude of the blue vector.

As mentioned above, the squared magnitudes can be used to calculate probabilities, but in practice the probabilities were so similar in higher-dimensional spaces as to be identical. There remains the possibility, however, that the magnitudes themselves remain sufficiently informative; this was borne out in practice.

3.2 Calculating the magnitude ratios

We have high-dimensional vectors, trained monolingually, of German and English sentences (§2.1). We consider their joint distribution by simply concatenating their vectors; there is no additional utility here in learning a translation between the monolingual spaces. We recenter the distribution to have zero mean—this simply makes the calculation and presentation easier—and transform the resulting matrix by $\Sigma^{-\frac{1}{2}}$.

For each sentence vector pair $\langle l_1, l_2\rangle$ (after recentering), we consider three vectors in the transformed space:

- the vector e_1 corresponding only to l_1's contribution to the concatenated and transformed vector (as if $l_2 = \vec{0}$)
- the vector e_2 corresponding only to l_2's contribution (as if $l_1 = \vec{0}$)
- the vector e corresponding to the transformation of the concatenation of l_1 and l_2

$$\begin{aligned} e_1 &= \Sigma^{-\frac{1}{2}}(l_1, \vec{0}) && (5)\\ e_2 &= \Sigma^{-\frac{1}{2}}(\vec{0}, l_2) && (6)\\ e &= \Sigma^{-\frac{1}{2}}(l_1, l_2) = e_1 + e_2 && (7) \end{aligned}$$

The measurement m we are interested in is the squared magnitude of the combined vector, divided by the sum of the squared magnitudes of e_1 and e_2 alone.

$$m = \frac{\|e\|_2^2}{\|e_1\|_2^2 + \|e_2\|_2^2} \tag{8}$$

Roughly speaking, does the sentence pair vector e in Mahalanobis space give more information (expressed in terms of its squared magnitude) than the component sentence vectors e_1 and e_2 do on their own? If so, we consider them unlikely to

p	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Mahalanobis	**0.977**	**0.976**	**0.974**	**0.972**	**0.972**
Linear	0.944	0.930	0.920	0.914	0.913
Nonlinear	0.871	0.871	0.897	0.900	0.905

Table 1: Accuracy of distinguishing parallel (i.e., related by a translation matrix T) vs. non-parallel (i.e., random) vectors, from a synthetic dataset of 100,000 pairs of 50-dimensional vectors, plus standard normal additive noise. p represents the proportion of parallel pairs in the dataset.

σ	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**
Mahalanobis	**.974**	**.778**	**.665**	**.617**	**.597**
Linear	.920	.722	.640	.606	.592
Nonlinear	.897	.658	.600	.586	.582

Table 2: Accuracy of distinguishing parallel (i.e., related by a translation matrix T) vs. non-parallel (i.e., random) vectors, from a synthetic dataset of 100,000 pairs of 50-dimensional vectors and "true" proportion $p = 0.3$, with varying degrees of additive noise. σ represents the standard deviation of the additive noise added to each of L1 and L2.

be parallel. We take the resulting value m to be the ranking (with lower values being better) for the post-filtering step described in §2.4.

Implementation-wise, we do not actually have to concatenate l_2 or l_1 with zeros in order to calculate (5) and (6), we can just multiply l_1 and l_2 by the relevant sub-matrix of $\Sigma^{-\frac{1}{2}}$. It is also unnecessary to actually transform the vector corresponding to the concatenation of $\langle l_1, l_2 \rangle$; the result is just the element-wise sum of e_1 and e_2.

```
def mahalanobis_whitening(X):
  # inverse square root of covariance
  cov = np.cov(X, rowvar=False)
  inv_cov = np.linalg.inv(cov)
  L, V = np.linalg.eig(inv_cov)
  diag = np.diag(np.sqrt(L))
  return V.dot(diag).dot(V.T)

def ssq(X): # sum of squares
  return np.sum(X*X, axis=1)

def mahalanobis_ratio(L1, L2):
  L1 -= L1.mean(axis=0)
  L2 -= L2.mean(axis=0)
  L = np.concatenate([L1,L2], axis=1)
  whitener = mahalanobis_whitening(L)
  E1 = L1.dot(whitener[:L1.shape[1],:])
  E2 = L2.dot(whitener[L1.shape[1]:,:])
  return ssq(E1+E2) / (ssq(E1) + ssq(E2))
```

Figure 2: Sample implementation of the Mahalanobis ratio calculation in Python, for two $n \times d$ NumPy arrays representing n samples of d-dimensional sentence vectors for two languages.

In code, this is a very simple calculation (only about 15 lines of Python+NumPy) and efficient (taking only a few minutes for millions of sentences), provided one has enough system memory to calculate it in one fell swoop. A sample implementation is given in Figure 2.

4 Internal results

4.1 Synthetic data

The unsupervised measurements on the sentence vectors were first tested on purely synthetic data: two sets of random normal vectors L1 and L2, in which some proportion p of vectors in L1 corresponded to L2 via a linear transformation T, and some proportion of vectors did not. We also added some Gaussian noise to each of L1 and L2, so that this transformation would not be perfect (as it would not be in real data). We varied the proportion of "true" pairs, and the proportion of additive noise, to test how robust these measurements would be in a variety of noise conditions.

Accuracy measurements on this data were made by thresholding scores so that the top p scores are set to 1.0 and the rest to 0.0.[5] This is also how we evaluate accuracy during sanity checking, below.

Table 1 contrasts three systems:

[5] Since the overall task is a *ranking* task, rather than a classification task, we do not at any point have to set a particular threshold for keeping data; this is a way in which the task at hand is easier than a typical anomaly detection task. We therefore simply use the correct proportion to set the thresholds.

1. (**Mahalanobis**) We perform the Mahalanobis ratio calculation described in §3.2.
2. (**Linear**) We learn a linear regression between L1 and L2, transform L1 according the resulting matrix, and measure the cosine similarity between the result and L2.
3. (**Nonlinear**) System (2), but instead of a linear regression we construct a simple two-layer perceptron with a ReLU nonlinearity.[6]

In each condition, the Mahalanobis measurement outperformed the other measurements. It may, of course, be that the conditions of this synthetic data are unlike real data—the relationship between the German and English sentence vectors might, for example, be better approximated with a nonlinear relationship—but, given the comparatively robust performance of the Mahalanobis measurement against a variety of noise conditions, we prioritized our development time to exploring it further.

4.2 Sanity checking

We also annotated about 300 random sentence pairs from the target corpus, according to whether we judged them to be parallel or not. We did not tune any parameters to this set, except to make sure that one hyperparameter, the dimensionality of the sentence vectors, did not lead to a numerical underflow condition as dimensionality increased.

Many of our initial attempts at measuring probabilities (and log probabilities) of sentence draws in higher dimensions (e.g. higher than 50) led to the differences between probabilities being so small that they could not be distinguished by floating-point representations, leading to a situation in which almost all probabilities were equivalent and no meaningful comparisons could be made, and thus to random performance when ranking sentences pairs. Keeping the measurements in terms of distances, and not converting them to probabilities, did appear to allow fine-grained comparison in higher dimensions, but we wanted to ensure that continuing to increase the dimensionality did not lead to indistinguishable measurements again.

Sanity checking (Table 3) confirmed that higher dimensionality does not necessarily lead to poorer discrimination: while 10-dimensional vectors only led to 44.1% accuracy in discriminating parallel from non-parallel pairs, 300-dimensional vectors gave 63.4% accuracy.

Dimensionality	10	50	100	300
Accuracy	.441	.548	.483	**.634**

Table 3: Sanity-checking results on 300 annotated sentences, for the Mahalanobis calculation (§3.2) on 10-, 50-, 100-, and 300-dimensional sentence vectors.

It is unclear why 100-dimensional vectors perform more poorly than both 50- and 300-dimensional vectors, but in any case this dataset only has 300 samples and we do not want to put too much stock in the results. The real purpose of this trial was to determine if the curse of dimensionality affects the Mahalanobis measurement adversely, and it does not appear to do so. We therefore used 300-dimensional vectors in our final submissions.

5 Official Results

Table 4 presents the results of the official evaluation, on seven corpora in four conditions. To help navigate the wall of numbers, keep in mind that we are mostly interested in the top unsupervised system `NRC-seve-bicov`, and that each table also presents average scores across the seven corpora, in the bottom right corner of each.

In the 100M-word conditions (that is to say, in the conditions where a statistical or neural machine translation system was trained on the top 100M words, as ranked by our filters), we find generally strong performance, with `NRC-seve-bicov` always performing above the median system and with most results in the top 10 (among 48 submissions).

However, we generally observe weaker downstream MT performance in 10M conditions, compared to other competitors; performing roughly near the median system in the NMT 10M condition and frequently below the median in the SMT 10M condition. This suggests to us that the unsupervised systems are adequate in finding

[6] We did not expect this to outperform the linear version—after all, there is no actual nonlinearity in the relationship between L1 and L2—but nonetheless wanted to see how a nonlinear regression would perform in different noise conditions. We observe, for example, that it does unsurprisingly poorly when only a low proportion p of sentences are related, a condition in which a linear regression performs comparatively well.

SMT, 10M-word								
	dev.	test						
domain	news	news	speech	laws	medical	news	IT	
corpus	newstest17	newstest18	iwslt17	Acquis	EMEA	GlobalVoices	KDE	average
top score	**23.23 (1)**	**29.59 (1)**	**22.16 (1)**	**21.45 (1)**	**28.70 (1)**	**22.67 (1)**	**25.51 (1)**	**24.58 (1)**
seve-bicov	19.66 (33)	25.96 (32)	18.64 (35)	18.78 (23)	**27.94 (5)**	20.05 (28)	21.38 (41)	22.13 (29)
mono-bicov	19.61 (35)	25.13 (36)	17.86 (39)	16.59 (35)	24.21 (37)	19.97 (34)	22.07 (37)	20.97 (38)
mono	17.98 (41)	23.49 (41)	16.63 (41)	15.49 (40)	23.09 (40)	18.65 (40)	21.39 (40)	19.79 (41)
SMT, 100M-word								
top score	**25.80 (1)**	**31.35 (1)**	**23.17 (1)**	**22.51 (1)**	**31.45 (1)**	**24.00 (1)**	**26.93 (1)**	**26.49 (1)**
seve-bicov	25.61 (11)	**31.11 (8)**	**22.84 (10)**	22.19 (15)	**31.20 (3)**	**23.67 (10)**	26.47 (18)	**26.25 (9)**
mono-bicov	**25.65 (5)**	**31.12 (5)**	**22.84 (10)**	**22.37 (8)**	**31.11 (7)**	**23.75 (7)**	26.19 (30)	**26.23 (10)**
mono	25.45 (14)	30.63 (21)	22.72 (20)	22.06 (21)	30.74 (20)	**23.70 (9)**	26.20 (28)	26.01 (19)

NMT, 10M-word								
	dev.	test						
domain	news	news	speech	laws	medical	news	IT	
corpus	newstest17	newstest18	iwslt17	Acquis	EMEA	GlobalVoices	KDE	average
top score	**29.44 (1)**	**36.04 (1)**	**25.64 (1)**	**25.57 (1)**	**32.72 (1)**	**26.72 (1)**	**28.25 (1)**	**28.62 (1)**
seve-bicov	24.49 (27)	30.32 (27)	21.47 (24)	22.57 (15)	**31.71 (2)**	23.08 (27)	22.89 (27)	25.34 (21)
mono-bicov	23.38 (30)	28.86 (32)	19.33 (34)	19.03 (29)	26.45 (32)	22.03 (32)	23.72 (23)	23.07 (30)
mono	20.83 (35)	24.97 (37)	17.19 (37)	16.57 (38)	23.79 (38)	19.75 (35)	21.85 (31)	20.69 (35)
NMT, 100M-word								
top score	**32.41 (1)**	**39.85 (1)**	**27.43 (1)**	**28.43 (1)**	**36.72 (1)**	**29.26 (1)**	**30.92 (1)**	**32.06 (1)**
seve-bicov	**32.10 (2)**	**39.39 (7)**	**27.09 (6)**	**28.31 (5)**	**36.30 (10)**	**28.94 (9)**	30.12 (16)	**31.69 (8)**
mono-bicov	**31.67 (9)**	38.86 (15)	**27.10 (5)**	**28.15 (9)**	35.96 (15)	28.87 (11)	30.41 (11)	31.56 (11)
mono	31.39 (16)	38.42 (21)	26.80 (12)	27.94 (12)	35.71 (21)	28.00 (27)	30.32 (14)	31.20 (19)

Table 4: BLEU scores (and ranking, out of 48 submissions) of NRC's unsupervised submissions: "seve" indicates single-feature (Mahalanobis ratio) parallelism assessment, "mono" indicates parallelism assessment using an unweighted ensemble of unsupervised features, "bicov" indicates that the final bigram coverage step (§2.4) was performed. Results in the top 10 performers are bolded.

a 100M word training set[7] but relatively poor at sub-selecting higher-quality sentences from that set. We think this may be because our system might have a bias towards picking relatively similar sentences, rather than the more diverse set of sentences that an MT training set needs, which is amplified in the 10M condition.

A surprising exception to this weakness is the European Medicines Agency (EMEA) corpus, in which `NRC-seve-bicov` is the #5 and #2 system in the SMT 10M and NMT 10M conditions, respectively. This could suggest that competitors are overfitting to the domain(s) of the training data, and performing correspondingly poorly on the out-of-domain EMEA, whereas `NRC-seve-bicov` cannot overfit in this manner. However, the other NRC unsupervised submissions, which also cannot overfit, have no special advantage on EMEA, and nor does `NRC-seve-bicov` perform notably well on other out-of-domain corpora in the 10M conditions.

6 Future research

The unsupervised methods described here seem promising in distinguishing parallel from non-parallel sentence pairs, but we interpret the 10M-word results as suggesting they are comparatively poor at distinguishing other MT-relevant features of sentence-pair quality. Considering bigram coverage (§2.4) appears to help somewhat, but more research is needed into mitigating the tendency of these measurements to prefer an uninteresting selection of sentences.

Also, it is likely that a sentence-vector, even a high-dimensional one, is not sufficiently fine-grained to choose the highest-quality pairs; the process described in this paper essentially says that two sentences with sufficiently similar topics are to be considered parallel, even if there is little word-level correlation between the sentences. We therefore intend to investigate a word-level analogue of the sentence-level Mahalanobis ratio measurement.

[7] Spot-checking a random sample of sentences suggested to us that there were indeed roughly 100M words worth of genuinely parallel data, but much of it would not have been particularly informative for machine translation. We therefore interpret 100M results as representing one's success at identifying parallel data, and the 10M results as representing how well one assesses usefulness-for-MT beyond parallelism.

References

Thomas Boggs. 2014. Whitening characteristics of the Mahalanobis distance. http://blog.bogatron.net/blog/2014/03/11/mahalanobis-whitening/.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit 2005*.

Philipp Koehn, Kenneth Heafield, Mikel L. Forcada, Miquel Esplà-Gomis, Sergio Ortiz-Rojas, Gema Ramírez Sánchez, Víctor M. Sánchez Cartagena, Barry Haddow, Marta Bañón, Marek Střelec, Anna Samiotou, and Amir Kamran. 2018a. ParaCrawl corpus version 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel Forcada. 2018b. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*.

Prasanta Chandra Mahalanobis. 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55.

Joel Martin, Howard Johnson, Benoît Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: Data driven machine translation and beyond, Volume 3*, pages 115–118. Association for Computational Linguistics.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540. Association for Computational Linguistics.

Irving S Reed and Xiaoli Yu. 1990. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(10):1760–1770.

Jörg Tiedemann. 2009. News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Rik Warren, Robert F Smith, and Anne K Cybenko. 2011. Use of Mahalanobis distance for detecting outliers and outlier clusters in markedly non-normal data: A vehicular traffic example. Technical report, SRA International Inc., Dayton, OH.