Lexical similarity can distinguish between automatic and manual translations

Agam Patel* and Dragomir R. Radev†,*

* Department of EECS
† School of Information
{agamrp,radev}@umich.edu
University of Michigan
Ann Arbor, MI 48109-1092

Abstract

We consider the problem of identifying automatic translations from manual translations of the same sentence. Using two different similarity metrics (BLEU and Levenshtein edit distance), we found out that automatic translations are closer to each other than they are to manual translations. We also use phylogenetic trees to provide a visual representation of the distances between pairs of individual sentences in a set of translations. The differences in lexical distance are statistically significant, both for Chinese to English and for Arabic to English translations.

1. Introduction

We try to compare different methods and see if automatic translations are more similar to each other or to manual translations. We use the edit distance between different translations of the Multiple-Translation Chinese Corpus (MTC) and the Multiple-Translation Arabic Corpus to compare the set of manual and automatic translations to each other. The MTC and the MTA corpus were developed by the Linguistic Data Consortium to support automatically evaluating translation quality. We chose at random 50 sentence sets from both the MTA and MTC corpus to run the experiments described in the rest of this paper.

We use BLEU to compare the same translation sentences and see if there is a correlation between BLEU scores and the edit distances in determining if automatic translations are similar to other automatic translations (and manual translations with other manual translations).

We use a hierarchical clustering method to visualize the distance between the sentences in the translation set by creating phylogenetic trees based on the score matrices from the Levenshtein edit distances and pairwise BLEU scores.

2. Related Work

We will discuss two different techniques using existing methods for translation evaluation to be used for translation comparison.

2.1. Levenshtein Edit Distance

The Levenshtein edit distance (Levenshtein, 1966) is a measure of the similarity between two strings. The distance is the number of deletions, insertions, or substitutions required to transform a sentence x into sentence y. The greater the edit distance is, the more different the two strings are. These are the distance values used to feed into the Fitch software to create the phylogenetic tree.

2.2. BLEU

BLEU (Papineni, 2002) is an automatic scoring method based on n-gram matching with reference translations. BLEU works by calculating the precision of unigrams up to n-grams between a test sentence and a reference sentence. To do this, BLEU counts the maximum number of times a word occurs in any single reference translation. The BLEU score ranges between 0 and 1, where higher scores are better. The score is calculated by taking the geometric mean of the precision scores. This is then multiplied by an exponential brevity penalty (which penalizes sentences that are too short).

2.3. Phylogenetic Trees and Fitch

Phylogenetic Trees and Fitch estimate phylogenies from the distance matrix data using the ”additive tree model” method according to which the distances are expected to equal the sums of branch lengths between the species (corresponding to different translations of the same sentence in our case). The ”additive tree model” is based on Fitch-Margoliash (Fitch, 1967) and least-squares distance methods. This method works by starting out with the two closest species and joining them under a node. It then determines the distance between the current tree and the rest of the tree. This is done until all the species have been added to the tree. The Fitch software takes an edit distance matrix as input and outputs a phylogenetic tree.

2.4. Multiple Sequence Alignment

Multiple sequence alignment techniques create a finite state automaton from aligning multiple translation of a sentence. This technique is primarily used for paraphrases and text generation, but it can also be used to MT evaluation. One technique is to align all of the reference translations and then compare each of them with the manual translation to get the edit distance.

Work has been done by Pang et al (Pang, 2003). and Barzilay and Lee (Barzilay, 2003) using multiple sequence alignment for paraphrase extraction and generation. Pang et al. also use the MTC corpus to create an FSA based on the alignment of the sentences. They hypothesize that their FSA can provide a good representation for MT evaluation comparable to BLEU. Barzilay and Lee use sequence alignment to create paraphrases represented by word lattice pairs to rewrite new sentences.

3. Experimental Setup

3.1. Methods

We used two methods when trying to compare the auto matic and manual translations. The first method involve using an edit distance matrix and BLEU scores to compar translations. The second method involved hierarchical clus tering using phylogenetic trees to show how similar transla tions were. In both cases the edit distance and BLEU scor are a measure of how similar/different two sentences ar For both methods, we picked a random translation sentenc set from the corpus.

For the first method, we compared the sentences and cre ated a Levenshtein edit distance matrix for each translatio set per sentence. We took a random sentence set from file in the MTC Corpus for a total of 50 sentence sets each wit 11 manual translations and 6 automatic translations. Th edit distance was calculated for each of these sentence se and a phylogenetic tree was produced using Fitch. Then for each of the 50 sentences, the order of their 11 manual translations and 6 automatic translations was randomized to produce another version of the edit distance matrix. The edit distance matrix has the following structure: $\left(\begin{smallmatrix}A & B\\ C & D\end{smallmatrix}\right)$, where A corresponds to the average edit distance of the manual translations to each other (the first 11 rows and 11 columns in the matrix). Region D corresponds to the average edit distance between the automatic translations (the last 6 rows and columns of the matrix). Regions B and C represent the average edit distance for the mix between manual and automatic translations.

The same method was used for the edit distance matrices for the randomly ordered sentences, except that now region A does not necessarily correspond to just manual translations and region D does not necessarily correspond to just automatic translations. Finally the A-D corresponds to the average distance of both region and A and D (manual and automatic translations) for an in-class comparison.

We also repeated this experiment using BLEU as the scoring metric. For this we did a mutual comparison of the sentences to create a score matrix. Using BLEU when comparing two sentences S1 and S2, it is possible to get different resulting scores depending on which sentence you designate as the test and reference. Using S1 as the reference sentence and S2 as the test sentence you will obtain a BLEU score. When you compare sentences S1 and S2 again but this time letting S2 be the reference sentence and S1 as the test sentence you will commonly obtain a different score than in the first case. In order to account for this issue and create a symmetric score matrix, we took the average value of the two scores obtained when comparing 2 sentences and used this as our score for the pair. Finally, in order to use the BLEU scores to build phylogenetic trees upon, we did a linear transformation on the scores ($1 - Score$) to obtain the final form of the BLEU score matrix. For example, in Table 2, the diagonal scores were originally 1 since comparing a sentence to itself yields a perfect score. After the transformation of the matrix, the diagonals are now 0 (the other values in the matrix have also been transformed accordingly). This transformation was necessary when using Fitch to build a phylogenetic tree based on BLEU scores because lower scores in the evolutionary tree indicate that sentences are more similar, which consequently correspond to higher BLEU scores.

其他党政及司法部门也必须从明年年初开始采取类似行动。

Figure 1: Chinese sentence.

In the second method, we created a phylogenetic tree based on the matrices for each translation set of a sentence to compare the similarity of the sentences. We built trees both by adding the translations to the trees in order (where the manual translations are added first and then the automatic translations) and also by adding the translations to the trees in random order. These phylogenetic trees provide a visual representation of similarity of the automatic and manual translations.

4. Data Sources Used

The data used for these experiments is the Multiple-Translation Chinese (MTC) Corpus and the Multiple-Translation Arabic (MTA) Corpus. The MTC was developed by the Linguistic Data Consortium to support automatically evaluating translation quality. It consists of a set of 11 manual translations and 6 automatic translations based on Mandarin Chinese sources. The Mandarin Chinese texts were taken from journal sources in the LDC and translated into English by both the manual and automatic translators. From these we chose random sets of sentences from the corpus to test. The data was modified slightly in order to be used with the various different software.

Similarly, the MTA corpus was also developed by the LDC to support automatically evaluation translation quality. The MTA consists of a set of 10 manual translations and 3 automatic translations for a an Arabic source sentence.

A sample of the actual sentence sets from the MTC corpus used for these experiments is shown in Figure 2. The Chinese source sentence that the translations in Figure 2 are based on is shown in Figure 1. Along with this sentence set, we ran the experiment on 49 more additional sentence sets in the MTC corpus.

5. Experimental Results

Shown in this section are the phylogenetic trees produced from the edit distances between the sentences in Figure 2. Table 1 shows the Levenshtein distances between the sentences. Table 2 shows the BLEU scores taken when comparing the sentences pairwise with each other. Figure 3 shows the resulting phylogenetic tree from the Levenshtein distance values on the left and the tree produced when built upon the BLEU scores on the right. Finally Table 3 shows the average of the Levenshtein distances for all 50 sentences when they are inserted in-order or random order. In these experiments, in-order corresponds to inserting the 11 manual sentences first and then the 6 automatic translations (i.e. exact order shown in Figure 2).

The experimental results for the phylogenetic trees and score matrices shown are from experiments on the Chinese translation to serve as representative examples. The exact

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17
S1	0	13	10	8	12	11	7	14	14	14	9	18	17	24	21	18	15
S2	13	0	13	11	11	9	11	7	11	13	12	20	18	24	18	21	13
S3	10	13	0	7	13	8	9	15	14	12	12	20	17	24	20	19	16
S4	8	11	7	0	9	8	7	13	12	12	7	18	16	24	20	17	14
S5	12	11	13	9	0	9	8	14	5	13	9	18	18	24	20	19	17
S6	11	9	8	8	9	0	7	12	9	9	8	17	16	24	20	17	13
S7	7	11	9	7	8	7	0	14	11	11	7	15	14	24	20	15	13
S8	14	7	15	13	14	12	14	0	14	17	15	21	19	23	19	22	16
S9	14	11	14	12	5	9	11	14	0	14	12	20	20	24	21	20	17
S10	14	13	12	12	13	9	11	17	14	0	11	20	21	24	21	20	19
S11	9	12	12	7	9	8	7	15	12	11	0	15	17	24	20	15	15
S12	18	20	20	18	18	17	15	21	20	20	15	0	16	22	20	7	13
S13	17	18	17	16	18	16	14	19	20	21	17	16	0	22	22	15	14
S14	24	24	24	24	24	24	24	23	24	24	24	22	22	0	21	24	22
S15	21	18	20	20	20	20	20	19	21	21	20	20	22	21	0	20	15
S16	18	21	19	17	19	17	15	22	20	20	15	7	15	24	20	0	12
S17	15	13	16	14	17	13	13	16	17	19	15	13	14	22	15	12	0

Table 1: Edit distance matrix for translations of the Chinese sentence.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17
S1	0.0000	1.0000	0.7839	0.7631	0.7909	0.8083	0.7763	1.0000	0.8380	0.8264	0.6429	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
S2	1.0000	0.0000	0.7714	1.0000	0.7732	0.5823	1.0000	0.4776	0.6777	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
S3	0.7839	0.7714	0.0000	0.5126	0.6705	0.4839	0.6662	1.0000	0.6928	0.7453	0.6326	1.0000	1.0000	1.0000	0.7429	0.7667	1.0000
S4	0.7631	1.0000	0.5126	0.0000	0.5621	0.6533	0.6841	1.0000	0.7190	0.7392	0.6257	1.0000	1.0000	1.0000	0.7175	0.7575	1.0000
S5	0.7909	0.7732	0.6705	0.5621	0.0000	0.5949	0.6289	1.0000	0.2647	0.6836	0.4957	1.0000	0.8069	1.0000	0.7669	0.7984	1.0000
S6	0.8083	0.5823	0.4839	0.6533	0.5949	0.0000	0.6015	0.6648	0.4796	0.6640	0.5870	1.0000	0.8075	1.0000	0.7838	0.8091	1.0000
S7	0.7763	1.0000	0.6662	0.6841	0.6289	0.6015	0.0000	1.0000	0.6661	0.6927	0.6243	1.0000	0.8183	1.0000	0.7965	0.8178	1.0000
S8	1.0000	0.4776	1.0000	1.0000	1.0000	0.6648	1.0000	0.0000	0.7466	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
S9	0.8380	0.6777	0.6928	0.7190	0.2647	0.4796	0.6661	0.7466	0.0000	0.6868	0.6696	1.0000	0.8176	1.0000	0.8079	0.8106	1.0000
S10	0.8264	1.0000	0.7453	0.7392	0.6836	0.6640	0.6927	1.0000	0.6868	0.0000	0.6640	1.0000	0.8328	1.0000	0.8120	0.8179	1.0000
S11	0.6429	1.0000	0.6326	0.6257	0.4957	0.5870	0.6243	1.0000	0.6696	0.6640	0.0000	1.0000	0.8042	1.0000	0.7072	0.7492	0.8120
S12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	0.8344	0.5492	1.0000
S13	1.0000	1.0000	1.0000	1.0000	0.8069	0.8075	0.8183	1.0000	0.8176	0.8328	0.8042	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000
S14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000
S15	1.0000	1.0000	0.7429	0.7175	0.7669	0.7838	0.7965	1.0000	0.8079	0.8120	0.7072	0.8344	1.0000	1.0000	0.0000	0.6884	0.7595
S16	1.0000	1.0000	0.7667	0.7575	0.7984	0.8091	0.8178	1.0000	0.8106	0.8179	0.7492	0.5492	1.0000	1.0000	0.6884	0.0000	0.8266
S17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8120	1.0000	1.0000	1.0000	0.7595	0.8266	0.0000

Table 2: Modified BLEU scores for Chinese sentence translations.

same experiments were run on the MTA corpus. The results (not shown here) from the experiments on the MTA corpus were very similar to the results when using the MTC corpus.

6. Analysis of Results

6.1. t-test analysis

The t-test results support our claim that Levenshtein edit distances and BLEU scores can be used as a metric for distinguishing automatic translations from manual translations. The test was performed by creating two populations from the score matrices used in the experiments. The first population consists of all the scores when comparing manual translations against the other manual translations along with the scores from comparing automatic translations against other automatic translations (regions A and D from the matrix as explained in the experimental methods section). The second population consists of the scores from regions B and C from the score matrices, which correspond to comparing automatic translations to manual translations. Specifically when looking at the sentences in the MTC corpus, the test was done by starting with the values from the 17×17 score matrices (i.e. Table 1 and Table 2). The first 11×11 values correspond to region A where there is a total of 121 separate values, and last 6×6 values corresponds to region D where there are a total of 36 values. All of these values were used as the first population data except for values that corresponds to a sentence compared to itself (i.e. the diagonals in the score matrix). So in total the first population data consisted of $(121 + 36 - 17) = 140$ separate values. The second population data consists of the 11×6 and 6×11 values, which corresponds to regions B and C for a total of $66 + 66 = 132$ separate values.

Table 4 shows the results of the t-tests performed using the above method on the Levenshtein edit distance and BLEU score matrices for both the translations sets of the MTC and MTA corpus. Both the Levenshtein edit distance and BLEU scores in the MTC corpus were adequate metrics to distinguish between the two populations. The same was true when working with the MTA corpus except for two sentences using BLEU scores. These sentences were particularly short so that the BLEU brevity penalty cause all the scores to be 0 or very close to 0 so that the sentences could not be distinguished. Overall the results from the test showed that the difference in scores are not likely to be due to chance and are really part of different populations.

	MTC (Chinese)		**MTA (Arabic)**	
Confidence value	**Edit**	**Bleu**	**Edit**	**Bleu**
$p > .05$	0	0	0	2
$1 * 10^{-10} < p < .05$	12	15	4	9
$p < 1 * 10^{-10}$	38	35	46	39

Table 4: Confidence values for Chinese and Arabic. This table shows the number of sentences out of 50 total that fell into each level of confidence, where sentences in the $p > .05$ category are not statistically significant.

```
  +----- S7
  !
  !     +----- S11
  !     !
  !     !           +----- S8
  !     !        +--6
  !     !        !  +----- S2
  !     !        !
  !     !        !        +----- [S17]
  !     !     +-10     +-15
  !     !     !  !     !  !  +----- [S16]
  !  +--2     !  !  +-12  +-14
  !  !  !     !  !  !  !     +----- [S12]
  !  !  !     !  !  !  !
  !  !  !     !  +-11  +----- [S13]
  !  !  !  +--4     !
  !  !  !  !  !     !  +----- [S15]
  !  !  !  !  !     +-13
  !  !  !  !  !        +----- [S14]
  !  !  !  !  !
--5--9  +--3  !  +----- S9
  !  !     !  +--7
  !  !     !     +----- S5
  !  !     !
  !  !     !  +----- S10
  !  !     +--8
  !  !        +----- S6
  !  !
  !  !  +----- S3
  !  +--1
  !     +----- S4
  !
  +----- S1
```

```
  +----- S11
  !
  !     +----- S10
  !  +--8
  !  !  +----- S7
  !  !
  !  !               +----- S8
  !  !            +--6
  !  !        +--10  +----- S2
  !  !        !   !
--9--5     +--2   +----- S6
  !  !     !  !
  !  !     !  !  +----- S9
  !  !  +--1  +--7
  !  !  !  !     +----- S5
  !  !  !  !
  !  !  !  !  +----- S4
  !  +--4  +--3
  !     !     +----- S3
  !     !
  !     !  +----- [S13]
  !     !  !
  !     +-11  +----- [S14]
  !        !  !
  !        +-12     +----- [S17]
  !           !  +-15
  !           !  !  +----- [S15]
  !           +-13
  !              !  +----- [S16]
  !              +-14
  !                 +----- [S12]
  !
  +----- S1
```

Figure 3: Fitch Trees of the Chinese sentence translations based on the *Levenshtein distances* (left) and *mutual BLEU scores* (right). Automatic translations are marked by square brackets, e.g., [S12].

6.2. Levenshtein vs. BLEU

Both Levenshtein edit distances and BLEU scores can be used to distinguish automatic from manual translations. As shown in Figure 3, trees based on both Levenshtein edit distance and BLEU scores separate into clusters of manual and automatic translation sentences. The difference between the two phylogenetic trees depends on which individual sentences are closer to each other. For example, the tree based on the Levenshtein edit distance shows that sentence 1 and 7 are close together while in the tree based on the BLEU scores, sentence 1 is close to sentence 11. This behavior is ubiquitous throughout the rest of the resulting phylogenetic trees produced for the sentence sets and is directly due to the different methods these two metrics apply in producing similarity scores.

Both metrics showed that the average score between automatic and manual translations were higher than when compared to the scores between automatic translations compared to other automatic translations and manual translations to other manual translations. Table 3 shows the average Levenshtein edit distance values while the BLEU scores are not shown in the paper, but have the same trend.

6.3. MTC vs. MTA

Conducting experiments Multiple-Translation Chinese and Multiple-Translation Arabic corpus yielded very similar results. All of the above experiments described in the paper which were run on the MTC corpus were also run on the MTA corpus. The similarity matrices for both Levenshtein edit distances and BLEU scores using the MTA corpus produced the same trends shown for the MTC corpus (i.e. scores between automatic and manual translations are higher than automatic against automatic or manual against manual translations). Similarly, from looking at the phylogenetic tree, automatic translations could be distinguished from manual translation since they separated into different clusters.

7. Conclusion

We presented two methods for using lexical similarity to help distinguish between automatic and manual translations using Chinese and Arabic corpora as examples. In both cases we used a hierarchical clustering method based on phylogenetic analysis (we should note that other hierarchical clustering algorithms would have most likely given us the same results).

The first method we used is to create similarity matrices between translations of the same sentence using two different metrics: Levenshtein edit distances and BLEU scores. The second method we employed was to create phylogenetic trees using the Fitch-Margoliash algorithm based on the score matrices produced in the first method. Using phylogenetic trees we show a division of automatic and manual translations based on a hierarchical clustering model.

Using Levenshtein edit distances and BLEU scores, we showed that the distance between manual and automatic translations on average is greater than the distance between manual translations compared to other manual translations (or automatic translations compared to other automatic translations). We showed that these differences in the scores (both Levenshtein edit distances and BLEU scores) are statistically significant. The results of the t-test performed strongly supports our claim that you can distinguish automatic from manual translations in the Multiple-

In-Order						Random Order					
SENT	A	B	C	D	A-D	SENT	A	B	C	D	A-D
1	14.28	17.92	17.92	14.28	14.28	1	14.99	17.36	17.36	13.94	14.75
2	22.26	33.80	33.80	26.83	23.31	2	26.05	30.47	30.47	26.33	26.11
3	28.43	34.24	34.24	22.83	27.15	3	30.76	31.77	31.77	24.06	29.22
4	9.26	16.94	16.94	14.61	10.48	4	13.01	14.14	14.14	12.28	12.84
5	38.81	53.80	53.80	40.33	39.16	5	41.32	50.35	50.35	44.56	42.06
6	20.41	35.03	35.03	32.33	23.15	6	27.40	30.36	30.36	25.94	27.07
7	5.75	9.00	9.00	7.78	6.22	7	6.41	8.32	8.32	8.06	6.79
8	43.69	56.41	56.41	43.39	43.62	8	47.06	52.67	52.67	45.78	46.76
9	29.69	42.18	42.18	35.17	30.94	9	31.97	39.68	39.68	36.67	33.04
10	33.85	48.98	48.98	38.00	34.80	10	38.26	44.47	44.47	39.72	38.60
11	31.50	44.23	44.23	33.22	31.90	11	38.20	39.03	39.03	29.78	36.27
12	23.32	36.30	36.30	32.17	25.35	12	29.75	31.88	31.88	26.78	29.07
13	12.50	20.52	20.52	18.22	13.81	13	17.12	17.41	17.41	14.06	16.42
14	52.94	63.35	63.35	49.67	52.19	14	49.54	64.11	64.11	58.33	51.55
15	9.26	17.64	17.64	15.72	10.74	15	13.70	14.61	14.61	11.89	13.29
16	17.29	22.35	22.35	18.33	17.53	16	20.68	20.73	20.73	12.89	18.89
17	41.21	53.33	53.33	43.33	41.69	17	46.99	49.42	49.42	38.22	44.98
18	7.26	11.80	11.80	9.11	7.68	18	9.17	10.08	10.08	9.00	9.13
19	36.20	54.30	54.30	46.11	38.47	19	42.86	48.89	48.89	43.56	43.02
20	12.00	16.15	16.15	13.22	12.28	20	13.21	15.02	15.02	13.33	13.24
21	38.74	57.11	57.11	47.72	40.80	21	43.62	52.68	52.68	47.56	44.52
22	23.70	34.64	34.64	28.94	24.90	22	30.50	30.20	30.20	22.39	28.64
23	9.87	18.92	18.92	14.72	10.98	23	12.02	16.59	16.59	16.06	12.94
24	11.31	17.61	17.61	15.17	12.19	24	14.99	15.02	15.02	12.28	14.37
25	21.44	34.53	34.53	28.72	23.11	25	26.41	30.58	30.58	26.50	26.43
26	27.29	35.86	35.86	27.06	27.24	26	28.68	34.05	34.05	29.06	28.76
27	8.66	10.62	10.62	8.67	8.66	27	8.96	10.36	10.36	8.61	8.88
28	24.21	32.52	32.52	23.22	23.99	28	27.09	29.39	29.39	25.00	26.61
29	10.64	13.32	13.32	10.00	10.50	29	11.60	12.42	12.42	10.06	11.25
30	27.77	40.33	40.33	32.11	28.76	30	32.18	36.30	36.30	32.06	32.15
31	10.69	14.15	14.15	11.33	10.84	31	12.28	12.97	12.97	10.33	11.83
32	52.66	67.70	67.70	51.17	52.32	32	55.19	63.88	63.88	56.67	55.53
33	22.94	38.05	38.05	31.33	24.87	33	28.94	33.05	33.05	29.50	29.07
34	13.88	17.42	17.42	12.50	13.57	34	14.64	16.27	16.27	14.17	14.54
35	24.91	36.27	36.27	28.22	25.67	35	26.63	33.86	33.86	31.28	27.69
36	13.82	19.92	19.92	16.61	14.46	36	16.28	18.00	18.00	15.39	16.08
37	30.36	41.44	41.44	29.83	30.24	37	34.00	37.58	37.58	31.78	33.49
38	18.86	31.91	31.91	27.39	20.82	38	23.40	28.17	28.17	25.83	23.96
39	18.15	27.12	27.12	20.78	18.75	39	21.59	23.83	23.83	21.28	21.52
40	17.29	22.35	22.35	18.33	17.53	40	20.23	20.38	20.38	15.67	19.18
41	31.04	44.23	44.23	34.06	31.73	41	35.57	40.03	40.03	34.22	35.26
42	24.31	32.09	32.09	25.89	24.68	42	25.92	30.44	30.44	26.56	26.06
43	15.01	20.82	20.82	17.50	15.58	43	16.99	19.12	19.12	17.06	17.01
44	6.84	8.76	8.76	6.72	6.82	44	7.54	8.12	8.12	6.72	7.35
45	25.60	35.56	35.56	28.94	26.37	45	30.89	32.06	32.06	24.00	29.31
46	5.69	9.41	9.41	8.67	6.37	46	7.47	8.14	8.14	7.33	7.44
47	34.74	47.17	47.17	36.61	35.17	47	39.92	42.67	42.67	35.72	38.96
48	6.03	8.70	8.70	7.22	6.31	48	6.43	8.11	8.11	8.06	6.80
49	22.13	28.06	28.06	17.72	21.12	49	23.47	25.71	25.71	21.83	23.10
50	30.45	47.58	47.58	39.00	32.41	50	40.41	40.86	40.86	30.11	38.05
Avg	22.38	31.65	31.65	25.22	23.03	Avg	25.65	28.83	28.83	24.57	25.40
StdDev	12.0	15.7	15.7	12.4	11.9	StdDev	12.7	14.8	14.8	13.1	12.7

Table 3: Average matrix score of edit distances for sentences with manual sentences as the first 11 and automatic translations as the last 6 compared with the random configuration. The column labeled A-D shows the average values for both regions A and D together.

Translation Chinese and Multiple-Translation Arabic corpus using lexical similarity methods.

8. Future Directions

Multiple sequence alignment can be used to align sentences and then running the test sentence in the same type of method to give an edit distance. Since the test sentence will be run against more sentences in essence (an aligned sen-

1. Other Party, governmental and law enforcement authorities must take similar actions beginning from the start of next year.

2.Other Party and government agencies and judicial departments must also take similar actions early next year.

3. All other Party, Government and Judicial Departments must start similar actions at the beginning of next year.

4. Other Party, government, and judicatory departments must take similar action at the beginning of next year.

5. Other party and government departments as well as judicial departments must take similar action from the beginning of next year.

6. All other party government and judicial departments must also take similar measures from the beginning of next year.

7. Other party and judicial authorities should take similar actions from the beginning of next year.

8. Other departments of the Party, the government and the judicial departments must also take similar actions early next year.

9. Other Party and Government departments as well as judicial departments must also take similar measures from the beginning of next year.

10. The other law enforcement agencies and departments will also take part in similar proceedings from the beginning of next year.

11. Other party, governmental and judicial departments will have to take similar action from the beginning of next year.

12. Other party politics and judicial department also will have to start from next year beginning of the year to adopt similar motion.

13. Other party s and judicial section must start from the beginning of year of next year taking similar action also .

14. The beginning of a year for and res judiciaria as welling must from next year of other party commences assuming is similar toing the proceeding.

15. At the beginning of next year politics and judicial department other parties must also start to pick to take similar action.

16. Other party politics and the judicial department also will have to start from at the beginning of next year to take the similar action.

17. Other party policies and judicial department must also begin from early next year to take similar action.

Figure 2: Sample sentence set from MTC corpora. Sentences 1-11 are manual translations, sentences 12-17 are automatic translations.

tence with multiple paths) this edit distance calculation may be a better metric to use in building the phylogenetic trees. The same experiments that were run on the Levenshtein edit distances and BLEU scores can then be extended to work on edit distances created by multiple sequence alignment.

Currently we can distinguish between automatic and manual translations by observing that they cluster into different groups by looking at the phylogenetic trees produced. We however cannot tell blindly which cluster of sentences is the manual translation or automatic translation. If we are given one already known manual or automatic translation, we can then extend the work to determine which cluster of sentences consists of manual translations and which is the automatic translation. This can be determined by observing where the known translation is added on the phylogenetic tree.

The method of using phylogenetic trees to visualize the similarity between sentences may also be extended to use in automatic translation evaluation to judge the quality of an automatic translation. When one knows which sentences are the manual translations, one can create a phylogenetic tree with both manual and automatic translations. The automatic translation that is the shortest distance away from the consensus sentence of the manual translations will likely have be the best quality automatic translation since it is the most similar to the set of manual translations. For example, when looking at tree on the right side of Figure 3 (the tree based on BLEU scores), assume that the node 4 is the consensus sentence of the manual translations. Since S13 is 15 units away from a node with manual translations, it is the closest automatic translation to the cluster of manual translations and likely to be the most similar to the manual translations.

9. Acknowledgments

This work was partially supported by the National Science Foundation under grant 0329043 “Probabilistic and link-based Methods for Exploiting Very Large Textual Repositories” administered through the IDM program. All opinions, findings, conclusions, and recommendations in this paper are made by the authors and do not necessarily reflect the views of the National Science Foundation.

10. References

K.A. Papineni, S. Roukos, T. Ward, W.J. Zhu, Bleu: a method for automatic evaluation of machine translation, In Proceedings of ACL-02 2002.

B. Pang, K. Knight, D. Marcu, Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences, HLT-NAACL 2003.

R. Barzilay, L. Lee, Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment, HLT-NAACL 2003

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215: 403-410.

http://www.ncbi.nlm.nigh.gov/BLAST/

W.M. Fitch and E. Margoliash, Construction of phylogenetic trees, Science, 155(1967), 279-284.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

V.I. Levenshtein. Binary codes capable of correcting insertions and reversals. Sov. Phys. Dokl., 10:707-10, 1966.

Michael Gilleland, Merriman Park Software, Levenshtein Distance, in Three Flavors. http://www.merriampark.com/ld.htm