Machine Translation Evaluation by Document Classification and Clustering

Feng He and Pascal Troemel
Swarthmore College,
500 College Avenue,
Swarthmore PA 19081, USA
{feng,troemel}@cs.swarthmore.edu

Abstract

We propose a Machine Translation evaluation system which does not require human-translated reference texts. The system makes use of a comparison between the performance of a computer's execution of NLP tasks on source text and on translated text to judge how effective the translation is. The difference in accuracy of the NLP task exectutions is used as a gauge for judging the competence of the Babelfish online translation system.

Keywords: *Machine Translation Evaluation, Document Classification, Clustering.*

1 Introduction

1.1 Machine Translation Evalutation

Machine translation research has been going on for several decades, and there are a number of systems available for use, mostly between English and a European or Asian language. Notable examples are products from Systran, which are used in Altavista's Babelfish online translation service. Machine translation evaluation has long been an extremely arduous task which requires much human input; more recently, the BLEU evaluation system [3] has made use of a much more automated, and thus more practical, approach. However, the BLEU system still requires the presence of several correct, human-translated reference texts (see Section 2.1 for an overview on the BLEU system). We propose a system which does not have this requirement, a system that is capable of judging the competence of a translation simply by comparing the source and target texts. We believe that this freedom from human input is important; human translation is a time-consuming and costly task in the MT evaluation process, and to cut it out alltogether will undoubtedly save resources.

We attempt to either prove or disprove the notion that although a machine translation may seem ineffective to a human reader, it still holds sufficient correct information to allow a computer to adequately perform the NLP tasks of text classification and clustering on it. If this is indeed the case, then even though machine translations may not yet be acceptable as accurate representations of works of literature in their original language, they may be far from useless to a computer capable of interpreting ("understanding") them.

Ultimately, a translation should be judged on how much information it retains from the original text. Following this notion, we judge the effectiveness of a translation system by comparing the accuracy results of a computer's NLP task execution on the source text and the target text. We expect a drop in performance that can then be interpreted as "acceptable" or "unacceptable," which serves as an evaluation of the system. Indeed the drop in performance gives us a quantitative measure of the translation's effectiveness.

In Section 2, we discuss a few relevant examples of previous work in the area of machine translation evaluation. Section 3 serves to describe how we collect data. The various experiments we performed are discussed in Section 4. In Sections 5 and 7 we share our results and conclusions.

1.2 Text Classification and Clustering

Text classification and clustering are two common NLP tasks that have been shown to obtain good results with statistical approaches. Classifications refers to the assigning of new documents to existing classes. The models for existing classes are built from documents known to belong to the classes. Usually a document is assigned to a single class, but it is also possible that a document has multiple class-tags.

Clustering refers to the dividing of a collection of documents into groups (clusters). These clusters are not pre-defined, although the number of clusters can be specified. It is also possible to create a hierarchy

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 35–39
Computer Science Department, Swarthmore College

of clusters, in which a cluster is further divided into sub-clusters.

Classification and clustering have long been studied, and there are many effective toolkits for both tasks. These two NLP tasks are natural choices for our experiments because they can be effectively performed on our data sets.

2 Related Work

2.1 MT Evaluation

The BLEU machine translation evaluation system [3] proposed in 2002 produced very respectable results, effectively emulating a human translation judge. The system produced a score for a given translation by comparing it to a group of "perfect" human-translated reference texts using n-gram precision values. After a few necessary details such as a brevity penalty had been added, the BLEU system's scores were found to correlate closely with those given by a variety of human judges on a group of test data. The main weakness of this system is its dependency on human-translated reference texts. Although it is far more automated than older, completely human-dependent "systems," which relied completely on human evaluation, the BLEU method still requires several correct translations. This means that, for every new text that the MT system is tested on, the BLEU evaluation system must first be presented with good reference texts, which must be produced by a group of humans. This can get expensive when a machine translation system is tested on numerous documents, a case that is clearly possible during the production of a truly effective MT system.

2.2 Tools

The Bow toolkit [1] was designed for statistical language modeling, text retrieval, classification and clustering. It provides a simple means of performing NLP tasks on a body of newsgroups, and was thus a very useful tool for us. We produced our results for text classification and clustering with the help of this system.

2.3 Other Related Works

In [4], Weiss et al. showed that newsgroup postings can be reasonably classified using statistical models.

Palmer et al. [2] investigated the effect of Chinese word segmentation on information retrieval. This work suggests that well-segmented Chinese text will improve performances of NLP tasks. Chinese segmentation is an active area of research, partly because current systems produce very poor segmentations. As we do not have a working segmenter for Chinese text, we expect our results to be accordingly affected.

Finally, Yang [5] gives a good overview of statistical text classification.

3 Data

The Internet is a rich resource for both English and Chinese texts. Chinese text on the Internet is encoded using Chinese character sets. There are several existing Chinese character sets. The most popular ones are: GB (simplied Chinese), Big5 (traditional Chinese) and UTF-8. Both GB and Big5 use two bytes to encode one character; UTF-8, a more recent encoding standard, uses three bytes. The differences between these encoding standards complicate data collection, as a data set must have uniform encoding to be usable. Character set detection and conversion are required. In addition, the character boundaries are often misaligned in online documents, producing corrupted Chinese text. These need to be removed from the data set.

For our experiments, we downloaded newsgroups articles as our data. Newsgroups are online communities where people send posts covering many different topics. There are newsgroups in both Chinese and English covering similar topics, providing parallel corpora. Postings are self-organized into topics, which serve as natural labels for document classification. The following data sets were downloaded:

- Data Set 1: Postings from 5 Chinese newsgroups were downloaded. The newsgroups and their topics are:
  - talk.politics.china – (Chinese politics)
  - cn.bbs.rec.movie – (movie reviews)
  - cn.sci.medicine – (Chinese medicine)
  - cn.culture.buddhism – (Buddhism)
  - cn.comp.software.shareware – (software/shareware)

  These newsgroups are chosen such that they are terminal groups (they are not further divided into subgroups), and they cover very different topics. About 800 postings were downloaded. The postings that contained corrupted Chinese or were too short (fewer than 50 Chinese characters) were removed, leaving about 400 to 700 usable postings from each newsgroup. The total number of postings is around 2000.

- Data Set 2: English translations of data set 1. We used Altavista's online Babelfish translation.

- Data Set 3: To create a parallel corpus to data set 1, we downloaded articles from 5 English newsgroups. The newsgroups are:

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 35–39
Computer Science Department, Swarthmore College

- talk.politics.usa – (American politics)
  - rec.arts.movies.reviews – (movie reviews)
  - misc.health.alternative – (alternative medicine)
  - soc.religion.christian.bible-study – (bible study)
  - comp.softawre.shareware.announce – (software/shareware)

  These newsgroups were chosen such that they cover similar topics as data set 1. About 3500 postings in all, roughly 700 from each group.

- Data Set 4: Chinese translations of data set 3 using Babelfish

4 Experiments

4.1 Experiment 1: Classification on Chinese Source

This experiment serves to compare the accuracy in performance of text classification on Data Sets 1 and 2: Chinese as source text and English as target text. We expected a significant drop in accuracy between the source and target performances, marking a loss of information in the translation. A typical member of Data Set 2, the target English, follows:

> *Perhaps in the very many movies, the audience already was used to it the good Lai shipyard -like violence and the love. But truly could attract the audience or has the male is positive just the charm actor, they usually could initiate audience's favorable impression even respect. Below is one good Lai shipyard cinema world first ten very male ranks announcement.*

Clearly the translated text is not good English, but it is also relatively clear that the topic of the posting is the movies, and that the correct newsgroup classification is *cn.bbs.rec.movie*, and not one of the other candidates: *cn.comp.software.shareware, cn.culture.buddhism, cn.sci.medicine,* or *talk.politics.china.* The purpose of this experiment is to discover whether a classification system is able to correctly classify documents such as this one.

We used the rainbow toolkit to classify the documents. To get a more rounded and reliable precision average for each data set, we performed classification using each of three different methods: Naive Bayes, TFIDF, and Probabilistic Indexing. Each data set was partitioned into training and testing sets. Either 80% or 60% of the documents from each class was randomly selected as traning data to build the class models. The remaining 20% or 40% was used as testing data and were classified. This procedure was repeated for 5 times each time different subsets was selected as training and testing data, and results averaged. The average results from data set 1 and 2 were compared. Note that the Chinese documents were not segmented, meaning each character was treated as a token, instead of a word, which usually consists of several characters. We expect this to lower classification performance on the Chinese documents.

4.2 Experiment 2: Classification on English Source

This experiment serves to compare the accuracy in performance of text classification on Data Sets 3 and 4: English as source text and Chinese as target text. Again, we expected a drop in accuracy between the source and target performances.

4.3 Experiment 3: Clustering on Chinese Source

In this experiment, we performed clustering on data sets 1 and 2 using *crossbow*, which is part of the rainbow toolkit. The number of clusters was specified to be 5, in accordance with the number of newsgroup topics. Note that, since no cluster topics were provided, the resulting clusters may or may not correspond to the original newsgroup topics. Indeed it is possible that articles from one newsgroup be divided into two clusters, while several newsgroups be merged into one cluster. However, there is usually a clear correspondence between the clusters and the original topics.

4.4 Experiment 4: Clustering on English Source

Experiment 4 is a repeat of experiment 3 on data sets 3 and 4.

5 Results

5.1 Experiment 1 and 2: Accuracy Results

	Classname	0	1	2	3	4	Total	Accuracy
0	software	318	19	1	22	.	360	88.33%
1	health	5	291	.	34	.	330	88.18%
2	movies	2	3	44	281	.	330	13.33%
3	religion	2	10	.	268	.	280	95.71%
4	politics	3	33	.	114	19	169	11.24%

Table 1. Classification with Probabilistic Indexing results example for Chinese as TARGET

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 35–39
Computer Science Department, Swarthmore College

Method	Test-Set Size 20%	40%
Naive Bayes	90.45	90.16
TFIDF	92.35	91.79
Probabilistic Indexing	86.85	86.14

Table 2. Classification accuracy, Chinese as SOURCE

Method	Test-Set Size 20%	40%
Naive Bayes	92.76	93.07
TFIDF	94.32	93.35
Probabilistic Indexing	90.40	90.39

Table 3. Classification accuracy, English as TARGET

Table 1 shows a typical classification result using documents from the English newsgroups. The rows represent the original newsgroup topics. The columns represent the number of documents assigned to each class. For example, of the 360 documents from *comp.software.shareware.announce*, 318 were assigned to class 0 (the correct class), 19 were assigned to class 1 *(mis.health.alternative)*, and so on.

Tables 2 and 3 summarize results from experiment 1. Each row records results using a specific modelling method. The size of the testing set was set to be either 20% or 40% of the total number of documents, and the results are tabled accordingly.

Method	Test-Set Size 20%	40%
Naive Bayes	97.38	97.74
TFIDF	97.60	97.97
Probabilistic Indexing	95.26	95.18

Table 4. Classification accuracy, English as SOURCE

Tables 3 and 4 summarize results from experiment 2, in which original English documents and their Chinese translations were classified.

5.2 Experiment 3 and 4: Clustering Results

Tables 6 and 7 summarize results from experiments 3 and 4. In each of the experiments, each data set was divided into 5 groups, which often corresponded to the original newsgroups. The clusters were matched with the newsgroups so that the total number of documents in wrong clusters was minimized. The second column in each table shows the number of correctly clustered documents out of the total number of documents. The third column gives the percentage accuracy.

Method	Test-Set Size 20%	40%
Naive Bayes	90.74	89.75
TFIDF	96.08	96.05
Probabilistic Indexing	65.42	65.38

Table 5. Classification accuracy, Chinese as TARGET

Texts	Accuracy	
Chinese as SOURCE	1385/1994	69.46%
English as TARGET	1431/1994	71.77%

Table 6. Clustering Accuracy

Texts	Accuracy	
English as SOURCE	1961/3674	53.38%
Chinese as TARGET	1817/3674	49.46%

Table 7. Clustering Accuracy

6 Discussion of Results

The results of Experiment 1 are unexpected: the target text actually performs better on the classification than the source text. This should obviously never occur, unless the machine translation system is so effective that it actually *corrects* errors instead of creating them. Since it is extremely unlikely that Babelfish is such a system, we need an alternate explanation. We propose two hypotheses, namely that either (1) the task of classifying Chinese text is somehow inherently more difficult than classifying English text, or (2) the lack of any segmentation in our Chinese mapping scheme is causing words to be incorrectly interpreted by the classification system. These two factors could easily be the reason that the results in Table 1 are actually lower than those of the target English in Table 2.

The results of Experiment 2, on the other hand, are more as expected. The source English performs much better than the target Chinese, as can be seen in Tables 3 and 4. These results suggest that the translation system did not perform very well; the accuracy average dropped from 96.68% to 82.86%, a 13.82% loss, which amounts to an error increase of about 500%. It is evident from Table 4 that the target Chinese results suffer greatly from the tests performed with the Probabilistic Indexing method. It is likely that information somehow key to this particular method was lost in the translation, and that this loss greatly hampered the classification of the documents. Table 5 shows

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 35–39
Computer Science Department, Swarthmore College

the results of a typical classification test-run using the PI method, and it is very interesting to note that the great majority of postings in the "movies" and "politics" newsgroups were incorrectly placed into the "religion" newsgroup. Reasons for this misplacement could be any of several possibilities, such as the all-encompassing nature of religion.

The same trend was observed in Experiments 3 and 4. Clustering result improved from Chinese documents to English translations, but deteriorated from English documents to Chinese translations. One interesting observation is that, clustering performed better on data sets 1 and 2 than 3 and 4. One possible reason is that data sets 3 and 4 are a lot bigger (contain almost twice as many documents).

It is difficult, however, to come up with a concrete measure of "acceptability" from these numbers. How do we know what is an acceptable drop in accuracy, or an unacceptable error increase? The answer to these questions may depend on the specific purpose of the MT system under evaluation: if its purpose is simply to provide a general idea of the original text, perhaps a 13.82% drop in accuracy is a perfectly adequate performance; but if its purpose is to provide an accurate, well-organized text fit for publication (which may be the purpose of future MT systems), a drop of even 2% may be unacceptable.

7 Conclusion and Future Directions

In this paper we proposed a machine translation evaluation system that is not bound to human-translated reference texts, instead making use of a text classification and clustering performance comparison. We described our experiments in which we evaluated the Babelfish translation system on newsgroup postings. The results were mixed. The Chinese to English translation actually improved classification and clustering performances, while the English to Chinese translation lowered performances. We hypothesize that this is either because Chinese text inherently does not fit well with the built-in language models in the Bow toolkit, or that the lack of segmentation hampered performance.

There are some interesting extensions to the experiments described in this paper. It will be interesting to see how much segmentation will improve task performances on the Chinese documents. We could also compare performances from other NLP task such as information retrieval. Finally, given that there are many NLP packages for English, and relatively few for Chinese, it is of practical value to see if it is possible to combine NLP packages with some machine translation system to obtain NLP packages for other languages.

References

[1] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[2] D. D. Palmer and J. D. Burger. Chinese word segmentation and information retrieval. *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia*, 2002.

[4] S. A. Weiss, S. Kasif, and E. Brill. Text classification in usenet newsgroups: A progress report. *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, 1996.

[5] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, vol 1:pp 69–90, 1999.

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 35–39
Computer Science Department, Swarthmore College