Improving Persian-English Statistical Machine Translation: Experiments in Domain Adaptation

Mahsa Mohaghegh
Massey University
School of Engineering and Advanced Technology
Auckland, New Zealand
M.Mohaghegh@massey.ac.nz

Abdolhossein Sarrafzadeh
Unitec
Department of Computing
Auckland, New Zealand
Hsarrafzadeh@unitec.ac.nz

Tom Moir
AUT University
School of Engineering
Auckland, New Zealand
Tom.moir@aut.ac.nz

Abstract

This paper documents recent work carried out for PeEn-SMT, our Statistical Machine Translation system for translation between the English-Persian language pair. We give details of our previous SMT system, and present our current development of significantly larger corpora. We explain how recent tests using much larger corpora helped to evaluate problems in parallel corpus alignment, corpus content, and how matching the domains of PeEn-SMT's components affect translation output. We then focus on combining corpora and approaches to improve test data, showing details of experimental setup, together with a number of experiment results and comparisons between them. We show how one combination of corpora gave us a metric score outperforming Google Translate for the English-to-Persian translation. Finally, we outline areas of our intended future work, and how we plan to improve the performance of our system to achieve higher metric scores, and ultimately to provide accurate, reliable language translation.

1 Introduction

Machine Translation is one of the earliest areas of research in Natural Language Processing. Research work in this field dates as far back as the 1950's. Several different translation methods have been explored to date, the oldest and perhaps the simplest being rule-based translation, which is in reality transliteration, or translating each word in the source language with its equivalent counterpart in the target language. This method is very limited in the accuracy it can give. A method known as Statistical Machine Translation (SMT) seems to be the preferred approach of many industrial and academic research laboratories, due to its recent success (Lopez, 2008). Different evaluation metrics generally show SMT approaches to yield higher scores.

The SMT system itself is a phrase-based translation approach, and operates using a parallel or bilingual corpus – a huge database of corresponding sentences in two languages.

The system is programmed to employ statistics and probability to learn by example which translation of a word or phrase is most likely to be correct. For more accurate translation results, it is generally necessary to have a large parallel corpus of aligned phrases and sentences from the source and target languages.

Our work is focussed on implementing a SMT for the Persian-English language pair. SMT has only been employed in several experimental translation attempts for this language pair, and is still largely undeveloped. This is due to several difficulties specific to this particular language pair. Firstly, several characteristics of the Persian language cause issues with translation into English, and secondly, effective SMT systems generally rely on large amounts of parallel text to produce decent results, and there are no parallel corpora of appropriate size currently available for this language pair. These factors are prime reasons why there is a distinct shortage of research work aimed at SMT of this particular language pair.

This paper firstly gives a brief background to the Persian language, focusing on its differences to English, and how this affects translation between the two languages. Next, we give details of our PeEn-SMT system, how we developed and manipulated the data, and aligned our parallel corpora using a hybrid sentence aligning method. We give a brief overview of previous tests with the earlier

Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, , pages 9–15,
Chiang Mai, Thailand, November 8, 2011.

version of the system, and then show our latest experiments with a considerably larger corpus. We show how increasing the size of the bilingual corpus (training model), and using different sizes of monolingual data to build a language model affects the output of PeEn-SMT system. We focus on the aim for a general purpose translator, and whether or not the increase in corpora size will give accurate results. Next we show that with the PeEn-SMT system equipped with different language models and corpora sizes in different arrangements, different test results are presented. We explain that the improved result variations are due to two main factors: firstly, using an in-domain corpus even of smaller size than a mixed-domain corpus of larger scale; secondly, spending much focus on stringent alignment of the parallel corpus. We give an overview of the evaluation metrics used for our test results. Finally, we draw conclusions on our results, and detail our plan for future work.

2 Persian Language Characteristics

Persian is an Indo-European language, spoken mostly in Iran, but also parts of Afghanistan, India, Tajikistan, the United Arab Emirates, and also in large communities in the United States. Persian is also known as Farsi, or Parsi. These names are all interchangeable, and all refer to the one language.

The written Persian language uses an extended Arabic alphabet, and is written from right to left. There are numerous different regional dialects of the language in Iran, however nearly all writing is in standard Persian.

There are several grammatical characteristics in written Persian which differ to English. There is no use of articles in Persian, as the context shows where these would be present. There is no capital or lowercase letters, and symbols and abbreviations are rarely used.

The subject in a Persian sentence is not always placed at the beginning of the sentence as a separate word. Instead, it is denoted by the ending of the verb in that sentence. Adverbs are usually found before verbs, but may also appear in other locations in the sentence. In the case of adjectives, these usually proceed after the nouns they modify, unlike English where they are usually found before the nouns.

Persian is a morphologically rich language, with many characteristics not shared by other languages (Megerdoomian & Laboratory, 2000). This can present some complications when it is involved with translation into *any* other language, not only English.

As soon as Persian is involved with statistical machine translation, a number of difficulties are encountered. Firstly, statistical machine translation of the Persian language is only recently being exploited. Probably the largest difficulty encountered in this task is the fact that there is very limited data available in the form of bilingual corpora.

The best language to pair with Persian for machine translation is English, since this language is best supported by resources such as large corpora, language processing tools, and syntactic tree banks, not to mention it is the most widely used language online, and in the electronic world in general.

When compared to English however, Persian has many differing characteristics, some of which pose significantly difficult problems for the task of translation. Firstly, compared to English, the basic sentence structure is generally different in terms of syntax. In English, we usually find sentence structure in its most basic form following the pattern of "subject – verb – object", whereas in Persian it is usually "subject – object – verb". Secondly, spoken Persian differs significantly from its written form, being heavily colloquial, to a much greater degree than English is. Thirdly, many Persian words are spelled in a number of different ways, yet all being correct. This in particular poses trouble for translation, since if one version of the spelling is not found in a bilingual corpus, such a word may be incorrectly translated, or remain as an OOV (out of vocabulary) word. Any SMT system designed for this language pair needs to take these details into consideration, and specifics of the system developed to cater for these differences.

3 PeEn-SMT Compositions

3.1 SMT System Architecture

The goal of a statistical machine translation system is to produce a target sentence $\boldsymbol{e}$ from a source sentence $\boldsymbol{f}$. It is common practice today to use phrases as translation units (Koehn et al., 2003; Och and Ney 2003) in the log-linear frame in order to introduce several models explaining the translation process.

The SMT paradigm relies on the probabilities of source and target words to find the best translation. The statistical translation process is given as:

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e}|\mathbf{f})$$

$$= \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{\mathcal{A}} \Pr(\mathbf{e}, \mathcal{A}|\mathbf{f}) \quad (1)$$

$$\approx \underset{\mathbf{e}}{\operatorname{argmax}} \max_{\mathcal{A}} \Pr(\mathbf{e}, \mathcal{A}|\mathbf{f}) \quad (2)$$

In the above equations, ($\boldsymbol{\mathcal{A}}$) denotes the correspondence between source and target words, and is called an alignment.

The Pr(e, $\boldsymbol{\mathcal{A}}$ |f) probability is modeled by combination of feature functions, according to maximum entropy framework (Berger, Pietra, & Pietra, 1996)

$$\Pr(\mathbf{e}, \mathcal{A}|\mathbf{f}) \propto \exp \sum_i \lambda_i f_i(\mathbf{e}, \mathcal{A}|\mathbf{f}) \quad (3)$$

The translation process involves segmenting the source sentence into source phrases *f*; translating each source phrase into a target phrase *e*, and reordering these target phrases to yield the target sentence **e***. In this case a phrase is defined as a group of words that are to be translated (Koehn, Och, & Marcu, 2003; Och & Ney, 2003) A phrase table provides several scores that quantize the relevance of translating *f* to *e*.

The PeEn-SMT system is based on the Moses SMT toolkit, by (Koehn, et al., 2007). The decoder includes a log-linear model comprising a phrase-based translation model, language model, a lexicalized distortion model, and word and phrase penalties. The weights of the log-linear interpolation were optimized by means of MERT(Och & Ney, 2003). In addition, a 5-gram LM with Kneser-Ney (Kneser & Ney, 2002) smoothing and interpolation was built using the SRILM toolkit (Stolcke, 2002).

Our baseline English-Persian system was constructed as follows: first word alignments in both directions are calculated with the help of a hybrid sentence alignment method. This speeds up the process and improves the efficiency of GIZA++ (Och & Ney, 2000), removing certain errors that can appear with rare words. In addition, all the experiments in the next section were performed using a corpus in lowercase and tokenized conditions. For the final testing, statistics are reported on the tokenized and lower-cased corpora.

3.2 Data Development

For optimum operation, a statistical language model requires a significant amount of data that must be trained to obtain proper probabilities. We had several Persian monolingual corpora available completely adapted to news stories, originating from three different news sources – Hamshahri (AleAhmad, Amiri, Darrudi, Rahgozar, & Oroumchian, 2009), IRNA[1] and BBC Persian[2] – Hamshahri contains around 7.3 million sentences, IRNA has almost 5.6 million, and the BBC corpus contains 7,005 sentences.

It is currently common to use huge bilingual corpora with statistical machine translation. Certain common language pairs have many millions of sentences available. Unfortunately for Persian/English , there is a significant shortage of digitally stored bilingual texts, and finding a corpus of decent size is a critical problem.

One English-Persian parallel text corpus we obtained consisted of almost 100,000 sentence pairs of 1.6 million words, and was mostly from bilingual news websites. There were a number of different domains covered in the corpus, but the majority of the text was in literature, politics, culture and science. Figure.1 shows the corpus divided into separate domains**.** To the best of our knowledge, the only freely available corpus for the English-Persian language pair is the TEP corpus, which is a collection of movie subtitles consisting of almost 3 million sentences - 7.8 million words. These two corpora were concatenated together to form News Subtitle Persian English Corpus (NSPEC) a single corpus of 3,100,000 sentences for use in one test, and will also be used in the future for further experiments.

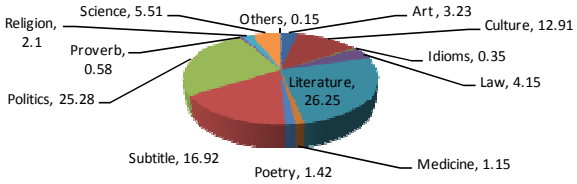

Figure 1. Domain percentages for NSPEC corpus

[1] http://www.irna.ir/ENIndex.htm
[2] http://www.bbc.co.uk/persian/

3.3 Alignment

The issue of word alignment in parallel corpora has been the subject of much attention. It has been shown that sentence-aligned parallel corpora are useful for the application of machine learning to machine translation, however unfortunately it is not usual for parallel corpora to originate in this form. The alignment of the corpus became a task of paramount importance, especially due to the shortage of bilingual text for English-Persian in the first place. There are several methods available to perform this task. Characteristics of an efficient sentence alignment method include speed, accuracy and also no need for prior knowledge of the corpus or the two languages. For the experiments presented in this paper, we used a hybrid sentence alignment method using sentence-length based and word-correspondence based models that covered all these areas, only requiring the corpus to be separated into word and sentence. In each of our experiments we firstly aligned the corpus manually using this hybrid method, and then later using GIZA++ when the data was put through Moses.

4 Experiments and Results

4.1 Overview of Previous Experiments

The original tests performed using PeEn-SMT as shown in some of previous papers produced unsatisfactory results (Mohaghegh, Sarrafzadeh, & Moir, 2010). It was initially thought that this was due to the small corpora and training models used. As detailed in these papers, a number of preliminary tests were carried out, and each time the language model was increased in size to a maximum of 7005 sentences. The training model at its largest consisted of 2343 sentences. The language model in these tests consisted of text collected from BBC news stories, and the training model consisted of a bilingual corpus of mostly UN news. It was thought that the unsatisfactory test results achieved could be remedied by enlarging the language model and corpus, since the amounts of data in each model were far too small to achieve any decent success in SMT.

4.2 Experiments

In order to develop the translation model, an English-Persian parallel corpus was built as explained in the Data Development section. We divided the parallel corpus into different sized groups for each test system. The details of the corpus size for each test are shown in Table 1. Table 2 shows the size of each test's corpus after the text was tokenized, converted to lowercase, and stripped of blank lines and their correspondences in the corpora. This data was obtained after applying the hybrid sentence alignment method.

Language Pair En-Pe	Data Genre	English Sentences	English words	Persian sentences	Persian Words
System1	Newswire	10874	227055	10095	238277
System2	Newswire	20121	353703	20615	364967
System3	Newswire	30593	465977	30993	482959
System 4	Newswire	40701	537336	41112	560276
System 5	Newswire	52922	785725	51313	836709
TEP	Subtitle	612086	3920549	612086	3810734
NSPEC	Newswire -Subtitle	678695	5596447	665678	5371799

Table 1: Bilingual Corpora Used to Train the Translation Model

Language Pair En-Pe	Data Genre	English Sentences	English Words	Persian sentences	Persian Words
System1	Newswire	9351	208961	9351	226759
System2	Newswire	18277	334440	18277	362326
System3	Newswire	27737	437871	27737	472679
System 4	Newswire	37560	506972	37560	548038
System 5	Newswire	46759	708801	46759	776154
TEP	Subtitles	612086	3920549	612086	3810734
NSPEC	Newswire Subtitle	618039	5370426	618039	5137925

Table 2: Bilingual Corpora after Hybrid Alignment Method

We divided the corpus to construct five different systems, beginning from 10,000 sentences in the smallest corpus, and increasing in steps of approximately 10,000 sentences each time up to the 5th test system, with a corpus of almost 53,000 sentences. In addition to the news stories corpus as shown earlier, we only had access to one freely available corpus, and this consisted of movie subtitles in Persian and English. This was shown to be in a completely different domain to our main corpus, so for most cases we preferred to run tests separately when using these corpora. Finally in NSPEC, we concatenated these two corpora, to ascertain the potential output with a combined corpus. We tested the subtitle corpus separately because we wished to see how an out-of-domain cor-

pus affected the result. In all cases, the test set consisted of a news article covering a variety of different domains showing various grammatical aspects of each language. In order to construct a language model, we used the transcriptions and news paper stories corpora. One source we used was the Hamshahri corpus, extracted from the Hamshahri newspaper, one of the most popular daily newspapers in Iran in publication for more than 20 years. Hamshahri corpus is a Persian text collection that consists of 700Mb of news text from 1996 to 2003. This corpus is basically designed for the classification task and contains more than 160,000 news articles on a variety of topics. Another source used was the IRNA corpus, consisting of almost 6 million sentences collected from IRNA (Islamic Republic News Agency). Table 3 summarizes the monolingual corpora used for the construction of the language model. SRILM toolkit (Stolcke, 2002)was used to create up to 5-gram language models using the mentioned resources. We tested the baseline PeEn-SMT system against different sizes of aligned corpora and different sized language models. Tables 4, 5 and 6 show the results obtained using the BBC, Hamshahri, and IRNA language models respectively.

Monolingual	Data Genre	Sentences	Words
BBC	News	7005	623953
Hamshahri (V.1)	News	7288643	65937456
IRNA	News	5852532	66331086

Table 3: Monolingual Corpora Used to Train the Language Model

4.3 Evaluation Metrics

One aspect of Machine Translation that poses a challenge is developing an effective automated metric for evaluating machine translation. This is because each output sentence has a number of acceptable translations. Most popular metrics yield scores primarily based on matching phrases in the translation produced by the system to those in several reference translations. The metric scores mostly differ in how they show reordering and synonyms.

In general, BLEU is the most popular metric used for both comparison of Translation systems and tuning of machine translation models (Papineni, Roukos, Ward, & Zhu, 2002); most systems are trained to optimize BLEU scoring. Many alternative metrics are also available however. In this paper we explore how optimizing a selection of different evaluation metrics effect the resulting model. The metrics we chose to work with were BLEU, IBM-BLEU, METEOR, NIST, and TER. While BLEU is a relatively simple metric, it has a number of shortcomings.

There have been several recent developments in evaluation metrics, such as TER (Translation Error Rate). TER operates by measuring the amount of editing that a human would have to undertake to produce a translation so that it forms an exact match with a reference translation (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006).METEOR (Denkowski & Lavie, 2010; Lavie & Denkowski, 2009) is a metric for evaluating translations with explicit ordering, and performs a more in-depth analysis of the translations under evaluation. The scores they yield tend to achieve a better correlation with human judgments than those given by BLEU (Snover, et al., 2006).

Another metric used was IBM-BLEU (Papineni, et al., 2002) , which performs case-insensitive matching of n-grams up to n=4.

BLEU and NIST (Zhang, Vogel, & Waibel, 2004) both produce models that are more robust than that of other metrics, and because of this, we still consider them the optimum choice for training.

4.4 Evaluation of the Results

Our first experiment was carried out with 10,000 sentences (System1) in the English-to-Persian translation direction. For comparison we tested the SMT model on different language models. As shown in Tables 4, 5, and 6, the best result was achieved when we trained the machine on the IRNA language model. We gradually increased the size of the corpora to the next test set (System 2), which was almost 21,000 sentences, and we repeated the test for different language models. Again the result showed that using IRNA resulted in the best translation, followed by BBC, then Hamshahri. We observed almost identical trends with each test set; up to the set with the largest corpus (53,000 sentences, System 5). It was originally thought that the dramatic increase in the size of both models would yield a much higher metric score, since it gave the translation program more data to work with. However, these new tests proved that this was not necessarily always true,

and corpus size alone was not synonymous with improved translation. For instance, in the case where the Hamshahri corpus was used for the language model, the output result was even worse than the original tests with a far smaller corpus like BBC. The IRNA corpus, larger than the original BBC corpus (7005 sentences) but still smaller than Hamshahri, yielded the best result of the two.

To establish a reason for the apparently illogical test results, the characteristics of each corpus were examined, together with their combinations in each test. After analysis, it was seen that there were a number of likely factors contributing to the poor results.

	Language Model =BBC news					
	Evaluation					
System	BLEU_4	MULTI_BLEU	IBM-BLEU	NIST	METEOR	TER
System 1	0.1417	10.96	0.0083	2.4803	0.3104	0.7500
System 2	0.1700	12.63	0.0172	2.5258	0.3347	0.6287
System 3	0.2385	24.66	0.0242	3.4394	0.3654	0.6312
System 4	0.2645	25.45	0.0274	3.6466	0.4466	0.6515
System 5	0.2865	26.88	0.0467	3.8441	0.4479	0.8181
TEP	0.1312	10.56	0.0095	2.6552	0.2372	0.8333
NSPEC	0.2152	19.94	0.0453	3.2643	0.3929	0.6824

Table 4: Automatic Evaluation Metrics of PeEn-SMT

	Language Model =Hamshahri					
	Evaluation					
System	BLEU_4	MULTI_BLEU	IBM-BLEU	NIST	METEOR	TER
System 1	0.1081	7.60	0.0246	2.1453	0.2526	0.8106
System 2	0.1229	8.77	0.0300	2.4721	0.3078	0.7196
System 3	0.1325	10.73	0.0149	1.2080	0.2215	0.7236
System 4	0.1945	10.87	0.0303	2.4804	0.2970	0.7500
System 5	0.2127	11.25	0.0288	3.6452	0.3040	0.8863
TEP	0.0127	1.05	0.0219	1.2547	0.1377	0.9015
NSPEC	0.0856	7.15	0.0499	1.9871	0.2313	0.7825

Table 5: Automatic Evaluation Metrics of PeEn-SMT System

	Language Model =IRNA					
	Evaluation					
System	BLEU_4	MULTI_BLEU	IBM-BLEU	NIST	METEOR	TER
System 1	0.2472	19.98	0.0256	3.5099	0.4106	0.6969
System 2	0.3287	29.47	0.0636	4.0985	0.4858	0.5833
System 3	0.3215	29.37	0.0565	4.1409	0.4838	0.5606
System 4	0.3401	30.99	0.0565	4.2090	0.4833	0.5833
System 5	***0.3496***	***29.25***	***0.0635***	***4.4925***	***0.5151***	***0.5236***
TEP	0.0535	3.98	0.0301	1.8830	0.2021	0.8787
NSPEC	0.1838	12.87	0.0366	3.0264	0.3380	0.7234

Table 6: Automatic Evaluation Metrics of PeEn-SMT System

One such factor involved the *nature* of the data comprising each corpus, and how this affected the match between the language model and the training model. For instance, in the case where we achieved an even lower score than the original tests, it was noted that the training model consisted of a bilingual corpus based mainly on *movie subtitles*, yet the Hamshahri corpus was a collection of *news stories*. For the most part, movies consist of spoken, natural language in everyday situations, filled with idioms, colloquial expressions and terms, and often incorrect grammar and sentence structure. These characteristics were heavily present in the training model. News stories on the other hand not only ideally consist of well-structured sentences, with correct grammar and little presence of colloquialism, but the very nature of this kind of literature is unique, and rarely found in natural language.

Another example showing this involved the subtitle corpus (TEP) that we had access to. This corpus was significantly larger in size (612,000 sentences) when compared to the other corpora that we had available to us. However, when we performed the same experiment against different language models, the result was quite unsatisfactory. We believe that this was due to our test sets being in a different domain than that of the movie subtitles.

These results led us to conclude that using larger language and training models alone was not a reliable determining factor in satisfactory output.

For the sake of comparison, Google Translator was tested on the same test data and results are in-

cluded in Tables 7. We compared our system to Google's SMT for this language pair, and compared to the evaluation metric score released by Google. Our PeEn-SMT system outperforms the Google translator in the English-to-Persian translation direction.

Google (English – Persian)						
System	BLEU_4	MULTI_BLEU	IBM-BLEU	NIST	METEOR	TER
Google	0.2611	21.46	0.0411	3.7803	0.5008	0.7272

Table 7: Automatic Evaluation Metric of Google Translator Output

5 Conclusion and Future Work

In this paper we presented the development of our English/Persian system PeEn-SMT. This system is actually a standard phrase-based SMT system based on the Moses decoder. The originality of our system lies mostly in the extraction of selected monolingual data for the language model. We used manual alignment of the parallel corpus, which was a hybrid sentence alignment method using both sentence length-based and word correspondence-based models, the results of which prove this method to be invaluable in obtaining a more accurate result from the system. We showed that increasing the size of the corpus alone cannot necessarily lead to better results. Instead, more attention must be given to the domain of the corpus. There is no doubt that the parallel corpora used in our experiments are small when compared to other corpora used in training SMT systems for other languages, such as German and Chinese, etc, or with Google, which has access to extensive resources. However we believe that the results from our system compare quite favorably, despite these shortcomings which we intend to address in our future work.

In the future we plan to develop a technique to find the most appropriate corpus and language model for PeEn-SMT system by detecting the domain of the input. We intend to perform tests using the matched-domain input, corpus and language models in an attempt to achieve even better translation results.

References

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems, 22*(5), 382-387.

Berger, A., Pietra, V., & Pietra, S. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics, 22*(1), 39-71.

Denkowski, M., & Lavie, A. (2010). Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages.

Kneser, R., & Ney, H. (2002). *Improved backing-off for m-gram language modeling*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). *Moses: Open source toolkit for statistical machine translation*.

Koehn, P., Och, F., & Marcu, D. (2003). *Statistical phrase-based translation*.

Lavie, A., & Denkowski, M. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation, 23*(2), 105-115.

Lopez, A. (2008). Statistical machine translation.

Megerdoomian, K., & Laboratory, N. M. S. U. C. R. (2000). *Persian Computational Morphology: A unification-based approach*: Computing Research Laboratory, New Mexico State University.

Mohaghegh, M., Sarrafzadeh, A., & Moir, T. (2010, 2010). *Improved Language Modeling for English-Persian Statistical Machine Translation.* Paper presented at the Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation,Coling, Beijing.

Och, F., & Ney, H. (2000). *Improved statistical alignment models*.

Och, F., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). *BLEU: a method for automatic evaluation of machine translation*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A study of translation edit rate with targeted human annotation*.

Stolcke, A. (2002). *SRILM-an extensible language modeling toolkit*.

Zhang, Y., Vogel, S., & Waibel, A. (2004). *Interpreting BLEU/NIST scores: How much improvement do we need to have a better system*.