Using the Productivity of Language is Rewarding for Small Data: Populating SMT Phrase Table by Analogy

Juan Luo*, **Aurélien Max****, **Yves Lepage***

* IPS, Waseda University, 808-0135 Fukuoka, Japan
** Univ. Paris Sud & LIMSI-CNRS, 91403 Orsay, France
juan.luo@suou.waseda.jp, aurelien.max@limsi.fr, yves.lepage@waseda.jp

Abstract

This paper is a partial report of the work on integrating proportional analogy into statistical machine translation systems. Here we present a preliminary investigation on the application of proportional analogy to generate translations of unseen n-grams from phrase table. We conduct experiments with different sizes of data and implement two methods to integrate n-gram pairs produced by proportional analogy into the state-of-the-art machine translation system Moses. The evaluation results show that n-grams generated by proportional analogy are rewarding for machine translation systems with small data.

Keywords: proportional analogy, phrase table, statistical machine translation, languages.

1. Introduction

Phrase-based machine translation model has made considerable advances in translation quality over the word-based model. Phrase-based statistical machine translation systems rely on parallel corpora for learning translation rules and phrases, which are stored in the so-called *phrase tables*. Thus, phrase table is the fundamental and vital component in the translation process. A phrase table is a list of phrase pairs that are translations of each other with feature scores. It is usually constructed in two steps: firstly, generating source-to-target and target-to-source word alignments; secondly, extracting bilingual phrase pairs from these alignments through heuristic combination of both directions.

Given a test sentence, those words that cannot be found in phrase table thus result in unknown words or phrases for a machine translation system. In this paper, we attempt to address the problem of unseen n-grams. Here, we propose to use proportional analogy to generate translations of unseen n-grams from phrase tables for statistical machine translation systems. We show that this method is useful for systems with small data. To the best of our knowledge, this paper is the first attempt at associating proportional analogy with phrase tables to generate strings that going beyond words, i.e., n-grams.

The remainder of this paper is organized as follows. Section 2 presents related work that inspired this study. In Section 3, we briefly introduce the proportional analogy. In Section 4, we describe our proposed method that is based on proportional analogy. In Section 5, experiments are reported and evaluation results are analyzed. Finally, we conclude in Section 6 with future works.

2. Related work

Several methods to deal with phrase tables in statistical machine translation systems have been proposed in the literature. Researches on trying to acquire additional data to increase translation coverage have focused on introducing paraphrases, n-grams, and multiword units.

In (Callison-Burch et al., 2006), paraphrases of unseen source phrases are incorporated into phrase tables to improve coverage and translation quality. However, their method is particularly pertinent to small corpus and out-of-vocabulary words. Augmenting phrase tables via paraphrasing is also investigated in (Marton et al., 2009; Fujita and Carpuat, 2013). A method of enlarging n-grams in phrase tables has been reported in (Ma et al., 2007), in which word packing is used to obtain 1-to-n alignments based on co-occurrence frequencies. They evaluated the performance on Chinese-to-English machine translation task and reported significant improvements. In (Henríquez Q. et al., 2010), collocation segmentation is performed on bilingual corpus to extract n-to-m alignments, which are used to augment phrase tables. However, the experimental results showed no difference in evaluation metric scores. Ren et al. (2009) proposed a strategy to extract domain bilingual multiword expressions and investigated methods to integrate these multiword units to phrase tables.

Proportional analogy has been researched and applied to address problems in various domains, for instance, transliteration, machine translation, handling unknown words, and so on.

In (Dandapat et al., 2010), methods of applying analogical learning over strings to the transliteration tasks were proposed. They showed that a combination of proportional analogy and statistical machine translation engine could lead to improvements over individual transliteration systems. Lepage and Denoual (2005) proposed an example-based machine translation system that is built upon proportional analogy. Their machine translation system works well on short sentences. Proportional analogy is also applied at the character-level to translate unknown words, which was reported in (Denoual, 2007) on Japanese-to-English task and (Langlais and Patry, 2007) on language pairs with close morphological structure. In (Langlais et al., 2009), they attempted to translate medical terms by using analogy and showed improvements.

Proportional analogy has been proposed to solve problems on the level of words, terms, and sentences. We are not aware of any previous work applying proportional analogy to strings between words and sentences, i.e., n-grams. This motivates the present paper. In this paper, we investigate whether proportional analogy can be used to generate unseen n-grams translations from phrase tables for statistical machine translation.

3. Proportional analogy

Proportional analogy is defined as a general relationship between four objects, i.e., four strings in this work. It is noted as $A : B :: C : D$, which is stated as "A is to B as C is to D". Analogy can be seen on the semantic or the formal levels. Here we work on the formal level only to the possible detriment of meaning.

Lepage (2004) proposed a formalization of analogies between strings. This formalization reduces to the counting of number of symbol occurrences and the computation of edit distances. The four strings, A, B, C and D, form an analogy only if:

$$A : B :: C : D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ \\ \delta(A,B) = \delta(C,D) \end{cases} \quad (1)$$

where $|A|_a$ stands for the number of occurrences of character a in string A. δ is the edit distance that involves only insertion and deletion with equal weights. $\delta(A,B)$ stands for the edit distance between strings A and B. As B and C may be exchanged in an analogy, the constraint on edit distance has also to be verified for $A : B :: C : D$, i.e., $\delta(A,C) = \delta(B,D)$. There is no need to verify the first constraint as, trivially, $|A|_a - |B|_a = |C|_a - |D|_a \Leftrightarrow |A|_a - |C|_a = |B|_a - |D|_a$.

As it is analyzed in (Lepage et al., 2007), proportional analogies can be written between words (2), chunks (3), or sentences (4):

$$\text{abundant} : \text{abundance} :: \text{present} : \text{presence} \quad (2)$$

$$\begin{matrix}\text{my room} \\ \text{key}\end{matrix} : \begin{matrix}\text{the room} \\ \text{key}\end{matrix} :: \begin{matrix}\text{my first} \\ \text{visit}\end{matrix} : \begin{matrix}\text{the first} \\ \text{visit}\end{matrix} \quad (3)$$

$$\begin{matrix}\text{Do you} \\ \text{like} \\ \text{music?}\end{matrix} : \begin{matrix}\text{Do you go} \\ \text{to conserts} \\ \text{often?}\end{matrix} :: \begin{matrix}\text{Do you like} \\ \text{classical} \\ \text{music?}\end{matrix} : \begin{matrix}\text{Do you go to} \\ \text{classical con-} \\ \text{certs often?}\end{matrix} \quad (4)$$

In this work, we focus on proportional analogies between sub-sentential strings, i.e., n-grams in phrase tables.

4. Unseen n-gram generation using analogy

In this section, we present our proposed method of generating entries by applying proportional analogical learning of unseen source n-grams in the test sentences. Instead of adding generated *analogy* n-gram as new entries to the baseline phrase table, we collect these entries to form an additional *analogy* phrase table.

The method comprises three stages:

(1) producing unseen n-grams;

(2) searching candidates in baseline phrase table;

(3) producing analogies of n-grams to form an *analogy* phrase table.

In the first stage, a test sentence is segmented into n-grams. A number of unseen n-grams (i.e., they are not found in baseline phrase table) are then extracted. In the second stage, given an unseen source n-gram, three candidate n-grams that should form analogical relationship with this unseen n-gram are searched in the source part of the baseline phrase table. After searching three source candidates, we thus obtain their corresponding n-grams in the target language. In the third stage, these three target n-grams are used to generate a new n-gram by proportional analogical learning. Finally, the newly generated n-gram pair is added as an entry to form an *analogy* phrase table.

4.1. An example

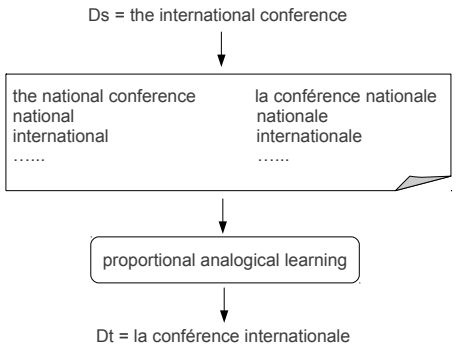

Figure 1: Example of analogical learning of n-grams from phrase table.

Let us illustrate the method clearer with an example (see Figure 1). Assume an unseen English source n-gram D_s in the test sentence:

$D_s = \textit{the international conference}$

To form an analogical relationship:

$A_s : B_s :: C_s : D_s$

Three candidates are searched in the source part of baseline phrase table and they are:

$A_s = \textit{international}$
$B_s = \textit{national}$
$C_s = \textit{the national conference}$

Their corresponding French target n-grams in phrase table thus are:

$A_t = \textit{internationale}$
$B_t = \textit{nationale}$
$C_t = \textit{la conférence nationale}$

A new n-gram D_t can be generated by proportional analogical learning:

$A_t : B_t :: C_t : D_t$
$D_t = \textit{la conférence internationale}$

Finally, we obtain a new n-gram pair (D_s, D_t), which can be added as an entry to *analogy* phrase table.

4.2. Feature scores

In the default phrase table of a standard statistical machine translation system, there are five feature scores: two translation probabilities and two lexical weights as proposed by Koehn et al. (2003), as well as the commonly used phrase penalty. Here, we calculate the feature scores of an *analogy* n-gram pair by two steps. In the first step, given three candidate n-gram pairs (A_s, A_t), (B_s, B_t), and (C_s, C_t), a feature score f of the *analogy* n-gram pair (D_s, D_t) is calculated by arithmetic mean as:

$$f(D_s, D_t) = \frac{1}{3} \sum_{x=A_i}^{C_i} f_x \tag{5}$$

For such an *analogy* n-gram pair, sets of analogies can be found in the phrase table. Thus, in the second step, it is calculated as:

$$f(D_s, D_t) = \frac{1}{n} \sum_{i=1}^{n} [\frac{1}{3} \sum_{x=A_i}^{C_i} f_x] \tag{6}$$

Here, in this preliminary investigation, we compute lexical weights as the above equations. We will compare the differences on the translation quality by using this computation and the one that originally proposed by Koehn et al. (2003) in the future.

5. Experiments

5.1. Experimental setup

In this section, we evaluate the performance of our proposed method empirically. Standard statistical machine translation systems were built by using the conventional pipeline: the Moses toolkit (Koehn et al., 2007), Batch MIRA (Cherry and Foster, 2012) to tune the parameters, the SRI Language Modeling (SRILM) toolkit (Stolcke, 2002) to build a 5-gram target language model with Kneser-Ney smoothing, and GIZA++ (Och and Ney, 2003) to generate word alignment. The maximum length of phrase pairs in phrase tables is set to 7 (the default phrase length in Moses).

The experiments were carried out using the Europarl parallel corpus (Koehn, 2005). We examined two language pairs: German-to-English and Polish-to-English. For each language pair, we tested two different sizes of training data. For the first setting, we used a training set of 347,614 and 350,000 sentence pairs, respectively. For the second setting, we used a small size dataset of 10,000 sentence pairs for both language pairs. We refer to the two settings as *Train=350k* and *Train=10k* in the following sections. The development set was made up of 500 sentence pairs, and test set contained 1,000 sentence pairs. A detailed description of the data sets is given in Table 1.

As for evaluation, four standard automatic evaluation metrics were used to assess the output of machine translation systems: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), WER (Nießen et al., 2000), and TER (Snover et al., 2006).

		German	English	Polish	English
train (350k)	sentences	347,614		350,000	
	tokens	9M	10M	8M	9M
	types	158,606	57,728	137,176	50,434
train (10k)	sentences	10,000		10,000	
	tokens	275,695	289,771	231,152	273,673
	types	23,626	13,344	24,424	10,390
dev.	sentences	500		500	
	tokens	14,062	14,697	13,294	14,989
	types	3,664	2,929	3,538	2,227
test	sentences	1,000		1,000	
	tokens	28,073	29,521	26,931	32,456
	types	5,888	4,381	4,856	2,791

Table 1: Statistics on the parallel corpus (M=million).

5.2. Experimental results

Since Moses supports multiple phrase tables, we investigated two methods to utilize *analogy* n-gram pairs: (1) Multiple PT, in which *either* phrase table is used for scoring; (2) backoff model, in which the second phrase table is used as a backoff for unknown words. We used *analogy* phrase table as backoff table and experimented on limiting the n-grams that were used from backoff table.

The results of experiments are shown in Table 2. Intuitively, *analogy* n-gram pairs are useful to improve the performance of statistical machine translation. However, from the results it can be seen that when given a training parallel corpus of approximately 350k sentences, the evaluation scores decrease slightly by comparing with the baseline. In the case of a small training data size (10,000 sentences), we can observe improvements over the baseline in both language pairs. The Multiple PT method achieves improvements in four evaluation metrics for Polish-English. However, this is not in consistent with the results obtained for language pair German-English, where decrease in translation quality is observed. The backoff model method improves translation quality in both languages. By limiting the phrase length in backoff table, i.e., *analogy* phrase table, the greatest increase in evaluation scores is obtained for 2-gram or 3-gram.

From the evaluation results in this table, we can conclude that n-gram pairs generated by proportional analogy are useful for translation systems with less training data.

5.3. Discussion

In order to examine n-gram pairs in detail, we analyzed the number of unique unseen n-grams in test sentences and those n-grams translations that can be generated by proportional analogy We also investigated the distribution of phrase lengths used during decoding.

An analysis of number of unique unseen n-grams in test set is shown in Table 3. From the table, it can be seen that the percentage of number of unique unseen n-grams translations that are generated by proportional analogy varies. The largest number of n-grams that can be produced by analogy are 2-gram and 3-gram in all cases.

Figure 2 shows phrase lengths that are actually used during the decoding process in all baseline systems. From the graph it can be seen that 50% to 80% of phrases are 1-to-1 translations. Further inspection shows that more than 90% of phrases used in decoding are of length up to 3. In order to know how the phrase length differs from the

		German-English				Polish-English			
		BLEU	NIST	WER	TER	BLEU	NIST	WER	TER
Train=350k	Baseline	**25.30**	**6.6193**	**55.27**	**59.65**	**34.26**	**7.5076**	**44.26**	**47.70**
	Multiple PT	24.68	6.4787	56.55	61.09	33.91	7.4041	45.01	48.56
	Backoff model (1gram)	25.16	6.5702	55.91	60.25	34.03	7.4790	44.29	47.73
	Backoff model (2gram)	24.90	6.5557	55.71	60.20	34.06	7.4563	44.46	47.94
	Backoff model (3gram)	24.96	6.5860	55.47	60.01	34.11	7.4695	44.60	48.07
	Backoff model (4gram)	24.90	6.5527	55.93	60.56	34.01	7.4519	44.61	48.16
	Backoff model (5gram)	24.95	6.5383	55.99	60.58	33.67	7.4202	44.92	48.34
	Backoff model (6gram)	24.48	6.5040	56.14	60.77	33.88	7.4162	44.78	48.43
	Backoff model (7gram)	24.54	6.4809	56.54	60.97	33.81	7.4264	44.80	48.42
Train=10k	Baseline	20.69	5.8867	58.76	63.73	19.35	5.6148	55.81	60.03
	Multiple PT	20.23	5.8173	59.61	64.87	**19.85**	**5.7568**	**55.64**	**59.91**
	Backoff model (1gram)	20.83	5.9339	59.10	63.95	19.94	5.8243	**54.84**	59.15
	Backoff model (2gram)	**20.92**	**5.9676**	58.76	**63.73**	19.95	**5.8305**	54.91	**59.11**
	Backoff model (3gram)	20.84	5.9205	**58.64**	63.80	**19.98**	5.8115	54.92	59.28
	Backoff model (4gram)	20.38	5.8712	58.89	64.11	19.76	5.7564	55.64	60.02
	Backoff model (5gram)	20.26	5.8104	59.51	64.79	19.71	5.7502	55.61	59.78
	Backoff model (6gram)	20.27	5.8211	59.68	64.67	19.68	5.7556	55.53	59.84
	Backoff model (7gram)	20.07	5.8149	59.65	64.93	19.74	5.7541	55.67	59.87

Table 2: Evaluation results.

		Train=350k		Train=10k	
	Total	Unseen	Analogy (%)	Unseen	Analogy (%)
German-English					
1-gram	5,888	1,516	879 (58%)	2,973	1,073 (36%)
2-gram	18,602	10,860	8,009 (74%)	14,946	9,073 (61%)
3-gram	24,001	20,014	14,426 (72%)	22,560	13,709 (61%)
4-gram	24,614	23,190	16,152 (70%)	24,142	13,704 (57%)
5-gram	23,926	23,359	15,726 (67%)	23,742	12,341 (52%)
6-gram	23,019	22,772	14,715 (65%)	22,943	10,846 (47%)
7-gram	22,064	21,927	13,612 (62%)	22,030	9,444 (43%)
Polish-English					
1-gram	4,856	402	307 (76%)	2,544	960 (38%)
2-gram	14,402	7,444	5,327 (72%)	12,708	5,990 (47%)
3-gram	18,034	14,780	9,633 (65%)	17,487	7,219 (41%)
4-gram	18,988	17,859	10,729 (60%)	18,865	6,413 (34%)
5-gram	18,864	18,528	10,066 (54%)	18,842	5,024 (27%)
6-gram	18,385	18,291	8,888 (49%)	18,379	3,719 (20%)
7-gram	17,782	17,761	7,599 (43%)	17,782	2,753 (15%)

Table 3: Number of unique n-grams in test set. Unseen: number of unique unseen n-grams. Analogy: number of unique unseen n-grams translations that are generated by proportional analogy.

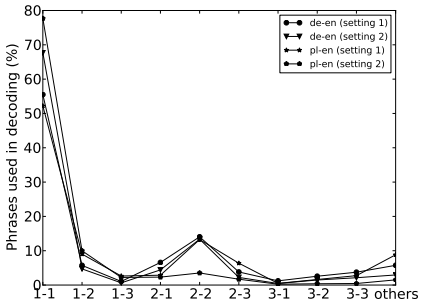

Figure 2: Distribution of phrases used during decoding (baseline systems).

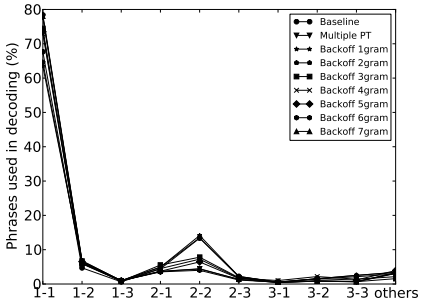

Figure 3: Distribution of phrases used during decoding (German-English; Train=10k).

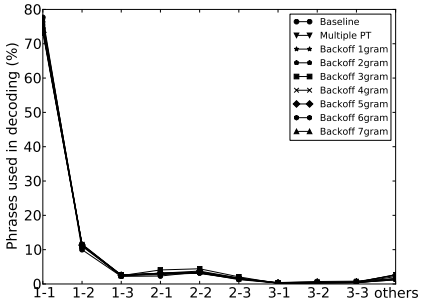

Figure 4: Distribution of phrases used during decoding (Polish-English; Train=10k).

baseline by using *analogy* n-gram pairs, we analyzed the distribution of phrases for all methods. The graphs are shown in Figure 3 and Figure 4. For German-English, there is a slight difference in the distribution of phrase lengths between the baseline and the method of using two phrase tables. For Polish-English, the distributions are approximately the same. In general, the majority of phrases used in decoding are up to 3-grams. To our disappointment, even though longer unseen n-gram translation pairs are generated by analogy, few of them are actually used.

6. Conclusion and future work

In this paper, we investigated the application of proportional analogy on generating unseen n-grams translations from phrase tables for statistical machine translation systems. We conducted experiments on two settings. The evaluation results reveal that populating phrase tables by proportional analogy is rewarding for machine translation systems with small amount of data.

Further inspections and experiments will be conducted in the future. We will analyze and exhibit which source phrases are actually better or worse translated with respect to the baseline (Max et al., 2010). We will also investi-

gate the issue of fabricating new phrase pairs (Chen et al., 2011). Experimental results show that the addition of more data is almost always beneficial, even though it may include inappropriate data, e.g., out-of-domain examples or noisy translations. This relates to the issue of identifying which phrases may benefit from additional data and special processing (Haddow and Koehn, 2012). Since the lexical weights of the *analogy* n-gram pairs are not computed in a conventional way in this paper, we would like to compare the effects on the translation quality by using different calculations in the future.

Acknowledgements

Part of the research presented in this paper has been done under a Japanese grant-in-aid (Kakenhi C, 23500187: Improvement of alignments and release of multilingual syntactic patterns for statistical and example-based machine translation). This research was partly supported by "Ambient SoC Global Program of Waseda University" of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Callison-Burch, C., Koehn, P. and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the NAACL-HLT*.

Chen, B., Kuhn, R. and Foster, G. (2011). Semantic smoothing and fabrication of phrase pairs for SMT. In *Proceedings of the 7th IWSLT*.

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the NAACL-HLT*.

Dandapat, S., Morrissey, S., Naskar, S. K. and Somers, H. (2010). Mitigating problems in analogy-based EBMT with SMT and vice versa: a case study with named entity transliteration. In *Proceedings of the 24th PACLIC*.

Denoual, E. (2007). Analogical translation of unknown words in a statistical machine translation framework. In *Proceedings of MT Summit XI*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of the 2nd HLT*.

Fujita, A. and Carpuat, M. (2013). FUN-NRC: Paraphrase-augmented phrase-based SMT systems for NTCIR-10 PatentMT. In *Proceedings of the 10th NTCIR*.

Haddow, B. and Koehn, P. (2012). Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the 7th WMT*.

Henríquez Q., A. C., Costa-jussà, R. M., Daudaravicius, V., Banchs, E. R. and Mariño, B. J. (2010). Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th ACL*.

Koehn, P., Och, F. J. and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the NAACL-HLT*.

Langlais, P. and Patry, A. (2007). Translating unknown words by analogical learning. In *Proceedings of EMNLP/CoNLL*.

Langlais, P., Yvon, F. and Zweigenbaum, P. (2009). Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Proceedings of the 12th EACL*.

Lepage, Y. (2004). Analogy and formal languages. *Electronic Notes in Theoretical Computer Science* 53:180–191.

Lepage, Y. and Denoual, E. (2005). ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages. In *Proceedings of the 2nd IWSLT*.

Lepage, Y., Migeot, J. and Guillerm, E. (2007). A measure of the number of true analogies between chunks in japanese. In *Proceedings of the 3rd LTC*.

Ma, Y., Stroppa, N. and Way, A. (2007). Bootstrapping word alignment via word packing. In *Proceedings of the 45th ACL*.

Marton, Y., Callison-Burch, C. and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the EMNLP*.

Max, A., Crego, J. M. and Yvon, F. (2010). Contrastive lexical evaluation of machine translation. In *Proceedings of the 7th LREC*.

Nießen, S., Och, F. J., Leusch, G. and Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of the 2nd LREC*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*.

Ren, Z., Lü, Y., Cao, J., Liu, Q. and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th AMTA*.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the 7th ICSLP*, volume 2.