Localization of Difficult-to-Translate Phrases

Behrang Mohit[1] and **Rebecca Hwa**[1,2]
Intelligent Systems Program[1]
Department of Computer Science[2]
University of Pittsburgh
Pittsburgh, PA 15260 U.S.A.
{behrang, hwa}@cs.pitt.edu

Abstract

This paper studies the impact that difficult-to-translate source-language phrases might have on the machine translation process. We formulate the notion of difficulty as a measurable quantity; we show that a classifier can be trained to predict whether a phrase might be difficult to translate; and we develop a framework that makes use of the classifier and external resources (such as human translators) to improve the overall translation quality. Through experimental work, we verify that by isolating difficult-to-translate phrases and processing them as special cases, their negative impact on the translation of the rest of the sentences can be reduced.

1 Introduction

For translators, not all source sentences are created equal. Some are straight-forward enough to be automatically translated by a machine, while others may stump even professional human translators. Similarly, within a single sentence there may be some phrases that are more difficult to translate than others. The focus of this paper is on identifying *Difficult-to-Translate Phrases* (DTPs) within a source sentence and determining their impact on the translation process. We investigate three questions: (1) how should we formalize the notion of difficulty as a measurable quantity over an appropriately defined phrasal unit? (2) To what level of accuracy can we automatically identify DTPs? (3) To what extent do DTPs affect an MT system's performance on other (not-as-difficult) parts of the sentence? Conversely, would knowing the correct translation for the DTPs improve the system's translation for the rest of the sentence?

In this work, we model difficulty as a measurement with respect to a particular MT system. We further assume that the degree of difficulty of a phrase is directly correlated with the quality of the translation produced by the MT system, which can be approximated using an automatic evaluation metric, such as BLEU (Papineni et al., 2002). Using this formulation of difficulty, we build a framework that augments an off-the-shelf phrase-based MT system with a DTP classifier that we developed. We explore the three questions in a set of experiments, using the framework as a testbed.

In the first experiment, we verify that our proposed difficulty measurement is sensible. The second experiment evaluates the classifier's accuracy in predicting whether a source phrase is a DTP. For that, we train a binary SVM classifier via a series of lexical and system dependent features. The third is an oracle study in which the DTPs are perfectly identified and human translations are obtained. These human-translated phrases are then used to constrain the MT system as it translates the rest of the sentence. We evaluate the translation quality of the entire sentence and also the parts that are not translated by humans. Finally, the framework is evaluated as a whole. Results from our experiments suggest that improved handling of DTPs will have a positive impact the overall MT output quality. Moreover, we find the SVM-trained DTP classifier to have a promising rate of accuracy, and that the incorporation of DTP information can improve the outputs of the underlying MT system. Specifically, we achieve an improvement of translation quality for non-difficult seg-

Proceedings of the Second Workshop on Statistical Machine Translation, pages 248–255,
Prague, June 2007. ©2007 Association for Computational Linguistics

ments of a sentence when the DTPs are translated by humans.

2 Motivation

There are several reasons for investigating ways to identify DTPs. For instance, it can help to find better training examples in an active learning framework; it can be used to coordinate outputs of multiple translation systems; or it can be used as means of error analysis for MT system development. It can also be used as a pre-processing step, an alternative to post-editing. For many languages, MT output requires post-translation editing that can be cumbersome task for low quality outputs, long sentences, complicated structures and idioms. Pre-translation might be viewed as a kind of preventive medicine; that is, a system might produce an overall better output if it were not thwarted by some small portion of the input. By identifying DTPs and passing those cases off to an expensive translation resource (e.g. humans) first, we might avoid problems further down the MT pipeline. Moreover, pre-translation might not always have to be performed by humans. What is considered difficult for one system might not be difficult for another system; thus, pre-translation might also be conducted using multiple MT systems.

3 Our Approach

Figure 1 presents the overall dataflow of our system. The input is a source sentence (a_1 ... a_n), from which DTP candidates are proposed. Because the DTPs will have to be translated by humans as independent units, we limit the set of possible phrases to be syntactically meaningful units. Therefore, the framework requires a source-language syntactic parser or chunker. In this paper, we parse the source sentence with an off-the-shelf syntactic parser (Bikel, 2002). From the parse tree produced for the source sentence, every constituent whose string span is between 25% and 75% of the full sentence length is considered a DTP candidate. Additionally we have a tree node depth constraint that requires the constituent to be at least two levels above the tree's yield and two levels below the root. These two constraints ensure that the extracted phrases have balanced lengths.

We apply the classifier on each candidate and select the one labeled as difficult with the highest classification score. Depending on the underlying classifier, the score can be in various formats such as class probablity, confidence measure, etc. In our SVM based classifier, the score is the distance from the margin.

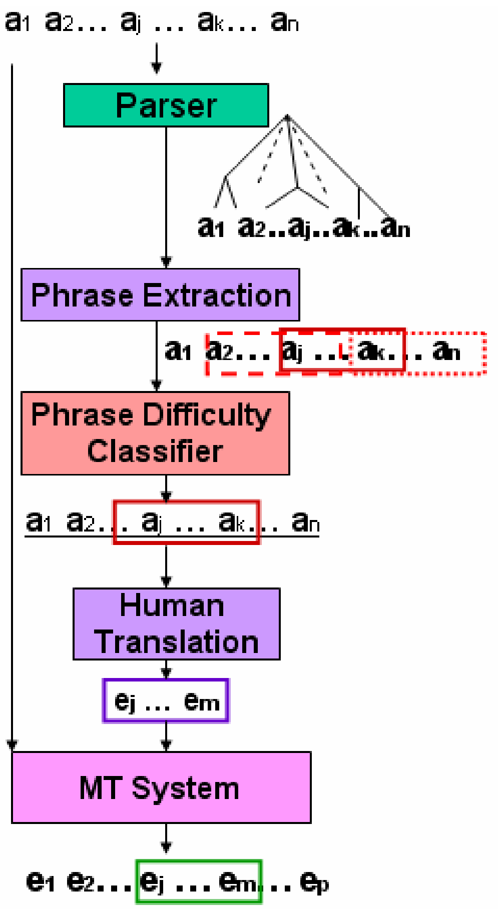

Figure 1: An overview of our translation framework.

The chosen phrase (a_j ... a_k) is translated by a human (e_i ... e_m). We constrain the underlying phrase-based MT system (Koehn, 2003) so that its decoding of the source sentence must contain the human translation for the DTP. In the following subsections, we describe how we develop the DTP classifier with machine learning techniques and how we constrain the underlying MT system with human translated DTPs.

3.1 Training the DTP Classifier

Given a phrase in the source language, the DTP classifier extracts a set of features from it and predicts whether it is *difficult* or not based on its feature values. We use an SVM classifier in this work. We train the SVM-Light implementation of the algorithm (Joachims 1999). To train the classifier, we need to tackle two challenges. First, we need to develop some appropriate training data because there is no corpus with annotated DTPs. Second, we need to determine a set of predictive features for the classifier.

Development of the Gold Standard

Unlike the typical SVM training scenario, labeled training examples of DTPs do not exist. Manual creation of such data requires deep understanding of the linguistics differences of source and target languages and also deep knowledge about the MT system and its training data. Such resources are not accessible to us. Instead, we construct the gold standard automatically. We make the strong assumption that difficulty is directly correlated to translation quality and that translation quality can be approximately measured by automatic metrics such as BLEU. We have two resource requirements – a sentence-aligned parallel corpus (different from the data used to train the underlying MT system), and a syntactic parser for the source language. The procedure for creating the gold standard data is as follows:

1. Each source sentence is parsed.
2. Phrase translations are extracted from the parallel corpus. Specifically, we generate word-alignments using GIZA++ (Och 2001) in both directions and combine them using the refined methodology (Och and Ney 2003), and then we applied Koehn's toolkit (2004) to extract parallel phrases. We have relaxed the length constraints of the toolkit to ensure the extraction of long phrases (as long as 16 words).
3. Parallel phrases whose source parts are not well-formed constituents are filtered out.
4. The source phrases are translated by the underlying MT system, and a baseline BLEU score is computed over this set of MT outputs.
5. To label each source phrase, we remove that phrase and its translation from the MT output and calculate the set's new BLEU score. If new-score is greater than the baseline score by some threshold value (a tunable parameter), we label the phrase as *difficult*, otherwise we label it as *not difficult*.

Rather than directly calculating the BLEU score for each phrase, we performed the round-robin procedure described in steps 4 and 5 because BLEU is not reliable for short phrases. BLEU is calculated as a geometric mean over n-gram matches with references, assigning a score of zero to an entire phrase if no higher-ordered n-gram matches were found against the references. However, some phrases with a score of 0 might have more matches in the lower-ordered n-grams than other phrases (and thus ought to be considered "easier"). A comparison of the relative changes in BLEU scores while holding out a phrase from the corpus gives us a more sensitive measurement than directly computing BLEU for each phrase.

Features

By analyzing the training corpus, we have found 18 features that are indicative of DTPs. Some phrase-level feature values are computed as an average of the feature values of the individual words. The following first four features use some probabilities that are collected from a parallel data and word alignments. Such a resource does not exist at the time of testing. Instead we use the history of the source words (estimated from the large parallel corpus) to predict the feature value.

(I) **Average probability of word alignment crossings**: word alignment crossings are indicative of word order differences and generally structural difference across two languages. We collect word alignment crossing statistics from the training corpus to estimate the crossing probability for each word in a new source phrase. For example the Arabic word *rhl* has 67% probability of alignment crossing (word movement across English). These probabilities are then averaged into one value for the entire phrase.

(II) **Average probability of translation ambiguity**: words that have multiple equally-likely translations contribute to translation ambiguity. For example a word that has 4 different translations with similar frequencies tends to be more ambiguous than a word that has one dominant translation. We collect statistics about the lexical translational ambiguities from the training corpus and lexical translation tables and use them to predict the ambiguity of each word in a new source phrase. The score for the phrase is the average of the scores for the individual words.

(III) **Average probability of POS tag changes**: Change of a word's POS tagging is an indication of deep structural differences between the source phrase and the target phrase. Using the POS tagging information for both sides of the training corpus, we learn the probability that each source word's POS gets changed after the translation. To overcome data sparseness, we only look at the collapsed version of POS tags on both sides of the corpus. The phrase's score is the average the individual word probabilities.

(IV) **Average probability of null alignments**: In many cases null alignments of the source words are indicative of the weakness of information about the word. This feature is similar to average ambiguity probability. The difference is that we use the probability of null alignments instead of lexical probabilities.

(V-IX) **Normalized number of unknown words**, **content words, numbers, punctuations:** For each of these features we normalize the count (e.g.: unknown words) with the length of the phrase. The normalization of the features helps the classifier to not have length preference for the phrases.

(X) **Number of proper nouns:** Named entities tend to create translation difficulty, due to their diversity of spellings and also domain differences. We use the number of proper nouns to estimate the occurrence of the named entities in the phrase.

(XI **Depth of the subtree**: The feature is used as a measure of syntactic complexity of the phrase. For example continuous right branching of the parse tree which adds to the depth of the subtree can be indicative of a complex or ambiguous structure that might be difficult to translate.

(XII) **Constituency type of the phrase**: We observe that the different types of constituents have varied effects on the translations of the phrase. For example prepositional phrases tend to belong to difficult phrases.

(XIII) **Constituency type of the parent phrase**

(XIV) **Constituency types of the children nodes of the phrase:** We form a set from the children nodes of the phrase (on the parse tree).

(XV) **Length of the phrase:** The feature is based on the number of the words in the phrase.

(XVI) **Proportional length of the phrase:** The proportion of the length of the phrase to the length of the sentence. As this proportion gets larger, the contextual effect on the translation of the phrase becomes less.

(XVII) **Distance from the start of the sentence and**: Phrases that are further away from the start of the sentence tend to not be translated as well due to compounding translational errors.

(XVIII) **Distance from a learned translation phrase:** The feature measure the number of words before reaching a learned phrase. In other words it s an indication of the level of error that is introduced in the early parts of the phrase translation.

3.2 Constraining the MT System

Once human translations have been obtained for the DTPs, we want the MT system to only consider output candidates that contain the human translations. The additional knowledge can be used by the phrase-based system without any code modification. Figure 2 shows the data-flow for this process. First, we append the pre-trained phrase-translation table with the DTPs and their human translations with a probability of 1.0. We also include the human translations for the DTPs as training data for the language model to ensure that the phrase vocabulary is familiar to the decoder and relax the phrase distortion parameter that the decoder can include all phrase translations with any length in the decoding. Thus, candidates that contain the human translations for the DTPs will score higher and be chosen by the decoder.

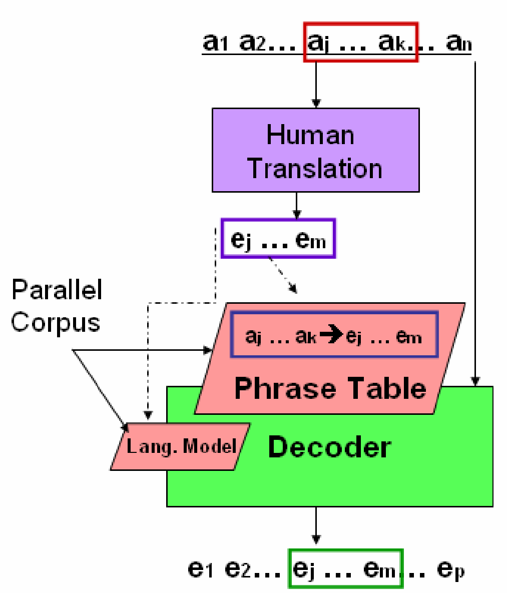
a_1 a_2... a_j ... a_k... a_n
Human Translation
e_j ... e_m
Parallel Corpus
a_j ... a_k → e_j ... e_m
Phrase Table
Lang. Model
Decoder
e_1 e_2... e_j ... e_m.. e_p

Figure 2: Human translations for the DTPs can be incorporated into the MT system's phrase table and language model.

4 Experiments

The goal of these four experiments is to gain a better understanding of the DTPs and their impact on the translation process. All our studies are conducted for Arabic-to-English MT. We formed a one-million word parallel text out of two corpora released by the Linguistic Data Consortium: Arabic News Translation Text Part 1 and Arabic English Parallel News Part 1. The majority of the data was used to train the underlying phrase-based MT system. We reserve 2000 sentences for development and experimentation. Half of these are used for the training and evaluation of the DTP classifier (Sections 4.1 and 4.2); the other half is used for translation experiments on the rest of the framework (Sections 4.3 and 4.4).

In both cases, translation phrases are extracted from the sentences and assigned "gold standard" labels according to the procedure described in Section 3.1. It is necessary to keep two separate datasets because the later experiments make use of the trained DTP classifier.

For the two translation experiments, we also face a practical obstacle: we do not have an army of human translators at our disposal to translate the identified phrases. To make the studies possible, we rely on a pre-translated parallel corpus to simulate the process of asking a human to translate a phrase. That is, we use the phrase extraction toolkit to find translation phrases corresponding to each DTP candidate (note that the data used for this experiment is separate from the main parallel corpus used to train the MT system, so the system has no knowledge about these translations).

4.1 Automatic Labeling of DTP

In this first experiment, we verify whether our method for creating positive and negative labeled examples of DTPs (as described in Section 3.1) is sound. Out of 2013 extracted phrases, we found 949 positive instances (DTPs) and 1064 negative instances. The difficult phrases have an average length of 8.8 words while the other phrases have an average length of 7.8 words[1]. We measured the BLEU scores for the MT outputs for both groups of phrases (Table 1).

Experiment	BLEU Score
DTPs	14.34
Non-DTPs	61.22

Table 1: Isolated Translation of the selected training phrases

The large gap between the translation qualities of the two phrase groups suggests that the DTPs are indeed much more "difficult" than the other phrases.

4.2 Evaluation of the DTP Classifier

We now perform a local evaluation of the trained DTP classifier for its classification accuracy. The classifier is trained as an SVM using a linear kernel. The "gold standard" phrases from the section 4.1 are split into three groups: 2013 instances are used as training data for the classifier; 100 instances are used for development (e.g., parameter tuning and feature engineering); and 200 instances are used as test instances. The test set has an equal number of difficult and non-difficult phrases (50% baseline accuracy).

In order to optimize the accuracy of classification, we used a development set for feature engineering and trying various SVM kernels and associated parameters. For the feature engineering part, we used the all-but-one heuristic to test the contribution of each individual feature. Table 2 presents the most and least contributing four features that we used in our classification. Among various features, we observed that the syntactic features are the most contributing sources of information for our classification.

Least Useful Features	Most Useful Features
Ft1: Align Crossing	Ft 2: Lexical Ambiguity
Ft 8: Count of Nums	Ft 11: Depth of subtree
Ft:9: Count of Puncs	Ft 12: Const type of Phr
Ft 10: Count of NNPs	Ft 13: Const type of Par

Table 2: The most and least useful features

The DTP classifier achieves an average accuracy of 71.5%, using 10 fold cross validation on the test set.

4.3 Study on the effect of DTPs

This experiment concentrates on the second half of the framework: that of constraining the MT system to use human-translations for the DTPs. Our objective is to assess to what degree do the DTPs negatively impact the MT process. We compare the MT outputs of two groups of sentences. Group I is made up of 242 sentences that contain the most difficult to translate phrases in the 1000 sentences we reserved for this study. Group II is a control group made up of 242 sentences with the least difficult to translate phrases. The DTPs make up about 9% of word counts in the above 484 sentences. We follow the procedure described in Section 3.1 to identify and score all the phrases; thus,

[1] Arabic words are tokenized and lemmatized by Diab's Arabic Toolset (Diab 2004).

this experiment can be considered an oracle study. We compare four scenarios:

1. **Adding phrase translations for Group I**: MT system is constrained using the method described in Section 3.2 to incorporate human translations of the pre-identified DTPs in Group I.[2]
2. **Adding phrase translations for Group II**: MT system is constrained to use human translations for the identified (non-difficult) phrases in Group II.
3. **Adding translations for random phrases**: randomly replace 242 phrases from either Group I or Group II.
4. **Adding translations for classifier labeled DTPs**: human translations for phrases that our trained classifier has identified as DTPs from both Group I and Group II.

All of the above scenarios are evaluated on a combined set of 484 sentences (group 1 + group 2). This set up normalizes the relative difficulty of each grouping.

If the DTPs negatively impact the MT process, we would expect to see a greater improvement when Group I phrases are translated by humans than when Group II phrases are translated by humans.

The baseline for the comparisons is to evaluate the outputs of the MT system without using any human translations. This results in a BLEU score of 24.0. When human translations are used, the BLEU score of the dataset increases, as shown in Table 3.

Experiment	BLEU
Baseline (no human trans)	24.0
w/ translated DTPs (Group I)	39.6
w/ translated non-DTPs (Group II)	33.7
w/ translated phrases (random)	35.1
w/ translated phrases (classifier)	37.0

Table 3: A comparison of BLEU scores for the entire set of sentences under the constraints of using human translations for different types of phrases.

While it is unsurprising that the inclusion of human translations increases the overall BLEU score, this comparison shows that the boost is sharper when more DTPs are translated. This is consistent with our conjecture that pre-translating difficult phrases may be helpful.

A more interesting question is whether the human translations still provide any benefit once we factor out their direct contributions to the increase in BLEU scores. To answer this question, we compute the BLEU scores for the outputs again, this time filtering out all 484 identified phrases from the evaluation. In other words in this experiment we focus on the part of the sentence that is not labeled and does include any human translations. Table 4 presents the results.

Experiment	BLEU
Baseline (no human trans)	23.0
w/ translated DTPs (Group I)	25.4
w/ translated non-DTPs (Group II)	23.9
w/ translated phrases (random)	24.5
w/ translated phrases (classifier)	25.1

Table 4: BLEU scores for the translation outputs excluding the 484 (DTP and non-DTP) phrases.

The largest gain (2.4 BLEU increment from baseline) occurs when all and only the DTPs were translated. In contrast, replacing phrases from Group II did not improve the BLEU score very much. These results suggest that better handling of DTPs will have a positive effect on the overall MT process. We also note that using our SVM-trained classifier to identify the DTPs, the constrained MT system's outputs obtained a BLEU score that is nearly as high as if a perfect classifier was used.

4.4 Full evaluation of the framework

This final experiment evaluates the complete framework as described in Section 3. The setup of this study is similar to that of the previous section. The main difference is that now, we rely on the classifier to predict which phrase would be the most difficult to translate and use human translations for those phrases.

Out of 1000 sentences, 356 have been identified to contain DTPs (that are in the phrase extraction list). In other words, only 356 sentences hold DTPs that we can find their human translations through phrase projection. For the remaining sentences, we do not use any human translation.

[2] In this study, because the sentences are from the training parallel corpus, we can extract human translations directly from the corpus.

Table 5 presents the increase in BLEU scores when human translations for the 356 DTPs are used. As expected the BLEU score increases, but the improvement is less dramatic than in the previous experiment because most sentences are unchanged.

Experiment	BLEU
Baseline (no human trans)	24.9
w/ human translations	29.0

Table 5: Entire Corpus level evaluation (1000 sentences) when replacing DTPs in the hit list

Table 6 summarizes the experimental results on the subset of the 356 sentences. The first two rows compare the translation quality at the sentence level (similar to Table 3); the next two rows compare the translation quality of the non-DTP parts (similar to Table 4). Rows 1 and 3 are conditions when we do not use human translation; and rows 2 and 4 are conditions when we replace DTPs with their associated human translations. The improvements of the BLEU score for the hit list are similar to the results we have previously seen.

Experiment on 356 sentences	BLEU
Baseline: full sent.	25.1
w/ human translation: full sent.	37.6
Baseline: discount DTPs	26.0
w/ human translation: discount DTPs	27.8

Table 6: Evaluation of the subset of 356 sentences: both for the full sentence and for non-DTP parts, with and without human translation replacement of DTPs.

5 Related Work

Our work is related to the problem of confidence estimation for MT (Blatz et. al. 2004; Zen and Ney 2006). The confidence measure is a score for n-grams generated by a decoder[3]. The measure is based on the features like lexical probabilities (word posterior), phrase translation probabilities, N-best translation hypothesis, etc. Our DTP classification differs from the confidence measuring in several aspects: one of the main purposes of our classification of DTPs is to optimize the usage of outside resources. To do so, we focus on classification of phrases which are syntactically meaningful, because those syntactic constituent units have less dependency to the whole sentence structure and can be translated independently. Our classification relies on syntactic features that are important source of information about the MT difficulty and also are useful for further error tracking (reasons behind the difficulty). Our classification is performed as a pre-translation step, so it does not rely on the output of the MT system for a test sentence; instead, it uses a parallel training corpus and the characteristics of the underlying MT system (e.g.: phrase translations, lexical probabilities).

Confidence measures have been used for error correction and interactive MT systems. Ueffing and Ney (2005) employed confidence measures within a trans-type-style interactive MT system. In their system, the MT system iteratively generates the translation and the human translator accepts a part of the proposed translation by typing one or more prefix characters. The system regenerates a new translation based on the human prefix input and word level confidence measures. In contrast, our proposed usage of human knowledge is for translation at the phrase level. We use syntactic restrictions to make the extracted phrases meaningful and easy to translate in isolation. In other words, by the usage of our framework trans-type systems can use human knowledge at the phrase level for the most difficult segments of a sentence. Additionally by the usage of our framework, the MT system performs the decoding task only once.

The idea of isolated phrase translation has been explored successfully in MT community. Koehn and Knight (2003) used isolated translation of NP and PP phrases and merge them with the phrase based MT system to translate the complete sentence. In our work, instead of focusing on specific type of phrases (NP or PP), we focus on isolated translation of difficult phrases with an aim to improve the translation quality of non-difficult segments too.

6 Conclusion and Future Work

We have presented an MT framework that makes use of additional information about difficult-to-translate source phrases. Our framework includes an SVM-based phrase classifier that finds the segment of a sentence that is most difficult to translate. Our classifier achieves a promising 71.5% accuracy. By asking external sources (such as human translators) to pre-translate these DTPs and using them to constrain the MT process, we im-

[3] Most of the confidence estimation measures are for unigrams (word level measures).

prove the system outputs for the other parts of the sentences.

We plan to extend this work in several directions. First, our framework can be augmented to include multiple MT systems. We expect different systems will have difficulties with different constructs, and thus they may support each other, and thus reducing the need to ask human translators for help with the difficult phrases. Second, our current metric for phrasal difficulty depends on BLEU. Considering the recent debates about the shortcomings of the BLEU score (Callison-Burch et. al. 2006), we are interested in applying alternative metrics such a Meteor (Banerjee and Lavie 2005). Third, we believe that there is more room for improvement and extension of our classification features. Specifically, we believe that our syntactic analysis of source sentences can be improved by including richer parsing features. Finally, the framework can also be used to diagnose recurring problems in the MT system. We are currently developing methods for improving the translation of the difficult phrases for the phrase-based MT system used in our experiments.

Acknowledgements

This work is supported by NSF Grant IIS-0612791. We would like to thank Alon Lavie, Mihai Rotaru and the NLP group at Pitt as well as the anonymous reviewers for their valuable comments.

References

Satanjeev Banerjee, Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pages 65–72*.

Daniel M. Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In Proceedings of *ARPA Workshop on Human Language Technology*

John Blatz, Erin Fitzgerald, George Foster, Simona Gan drabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Summer Workshop Final Report.

Chris Callison-Burch, Miles Osborne, and Philip Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In Proc. of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. *In Proceeding of NAACL-HLT 2004*. Boston, MA.

Thorsten Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas*, pages 115–124

Philipp Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proceedings of 41st the Annual Meeting on Association for Computational Linguistics (ACL-2003)*, pages 311–318.

Franz Och, 2001, “Giza++: Training of statistical translation model”: http://www.fjoch.com/GIZA++.html

Franz. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics, 29(1):19–51*.

Kishore Papineni and Salim Roukos and Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL-2002)*, Pages 311-318, Philadelphia, PA

Nicola Ueffing and Hermann Ney. 2005. Application of word-level confidence measures in translation. In *Proceedings of the conference of the European Association of Machine Translation (EAMT 2005)* , pages 262–270, Budapest, Hungary

Richard Zens and Hermann Ney, 2006. N -Gram Posterior Probabilities for Statistical Machine Translation. In Proceedings of ACL Workshop on Statistical Machine Translation. 2006