Semi-Automatic Error Analysis for Large-Scale Statistical Machine Translation Systems

Katrin Kirchhoff*, Owen Rambow†, Nizar Habash†, Mona Diab†

*Dept. of Electrical Engineering
University of Washington
Seattle, WA, 98105, USA
katrin@ee.washington.edu

†Center for Computational Learning Systems
Columbia University
475 Riverside Drive, New York, NY, 10115, USA
{rambow,habash,mdiab}@cs.columbia.edu

Abstract

.
This paper presents a general framework for semi-automatic error analysis in large-scale statistical machine translation (SMT) systems. The main objective is to relate characteristics of input documents (which can be either in text or audio form) to the system's overall translation performance and thus identify particularly problematic input characteristics (e.g. source, genre, dialect, etc.). Various measurements of these factors are extracted from the input, either automatically or by human annotation, and are related to translation performance scores by means of mutual information. We apply this analysis to a state-of-the-art large-scale SMT system operating on Chinese and Arabic text and audio documents, and demonstrate how the proposed error analysis can help identify system weaknesses.

Introduction

State-of-the-art large-scale statistical machine translation (SMT) systems are fairly complex: they typically consist of multiple component models (e.g. translation model, language model, reordering model), they perform multiple decoding passes, and have millions of parameters that may interact in non-transparent ways. At the same time, translation input is becoming increasingly varied, consisting not only of newstext-style documents or parliamentary proceedings but also of unstructured text sources such as emails, blogs, or newsgroup texts. As a consequence, diagnosing problems in translation performance and relating them to characteristics of the translation input is growing more and more difficult. Problems are aggravated further in the case of speech translation (of e.g. broadcast news, talkshows, etc.), where the input to the translation module is provided by an automatic speech recognition (ASR) system whose performance also influences the quality of the final translation output.

During machine translation (MT) system development, automatic evaluation criteria are commonly used to judge performance, such as the BLEU score (Papineni et al. 2002), the NIST score (Doddington 2002), or, more recently, METEOR (Banerjee & Lavie 2005). Although the use of fully automated evaluation criteria is helpful in accelerating the system development cycle, all of the above criteria have shown to be inferior to human judgments of translation performance. Moreover, they do not yield any insight into precisely which input characteristics caused particular translation errors, or which system components need to be improved in order to reach the desired performance level. Human analysis of machine translation errors, on the other hand, is costly and time-intensive and can typically not be performed on a regular basis in the course of system development. Thus, an automatic or semi-automatic procedure for better error analysis would be desirable. In this paper we present a semi-automatic error analysis procedure for large-scale MT systems that is designed to identify characteristics of input documents, as well as glitches in a multi-component pipelined system structure, that are responsible for poor translation output. Measurements of characteristics such as source, genre, style, dialect, etc. are extracted automatically or obtained from human annotations and are statistically related to measurements of the overall system performance. The various input document features are then ranked with respect to their impact on translation performance.

Previous Work

Most work on error analysis in statistical machine translation has made use of extensive human analysis, such as classifying unsatisfactory output into categories such as wrong word choice, missing content words, missing function words, etc. (see e.g. Koehn 2003, Och et al. 2003).

Previous work on automatic or semi-automatic error analysis in SMT systems includes Niessen et al. (2000), Popovic et al. (2006a) and Popovic et al. (2006b). In Niessen et al. (2002), a graphical user interface was presented that automatically extracts various error measures for translation candidates and thus facilitates manual error analysis. In Popovic et al. (2006a) and Popovic et al. (2006b), errors in an English-Spanish statistical MT system were analyzed with respect to their syntactic and morphological origin. This was done by modifying the references and the machine translation output by eliminating morphological inflections or suspected reordered constituents, and by analyzing the resulting changes in position-independent word error rate or word error rate. This revealed system problems in specific areas of inflectional morphology and syntactic reordering.

To our knowledge, there have not been any previous attempts at automatically relating a wide range of features of the translation input (e.g., document style, dialect, source, topic) to the output performance, which is the problem addressed here. Such an analysis is complementary to error analyses that are internal to the translation model: in highlighting features of the translation input, it can result in different ways of preprocessing or pre-classifying input documents, or, in cases where the input is the output from a different processing module, it can lead to better overall system integration. Improvements in these two areas will become increasingly important for the type of large-scale, complex translation systems that are beginning to be developed.

Data and System

This analysis was performed within the context of the 2006 machine translation evaluations of the US-based GALE project. Systems participating in the 2006 evaluation were expected to translate documents in two different languages (Arabic and Chinese) and four different genres: broadcast news (BN), broadcast conversations (BC), newswire text (NW) and newsgroups text (NG). The first two genres represent audio conditions, i.e. the documents are provided as waveform files and first need to be processed by an ASR module whose output then serves as the input to the machine translation component. The latter two are text conditions. In all cases, the target language is English. The MT system used for this study is a combination of the outputs of several individual SMT systems (developed by RWTH Aachen, SRI, NRC, and University of Washington, respectively). The combination was done as described in Matusov et al. (2006). MT performance was measured by standard scoring techniques such as BLEU, METEOR, etc., and, additionally, by human translation error rate (HTER, see Snover et al. (2005)). HTER is based on a comparison of an MT output hypothesis with human reference translations that were created specifically for this output by performing edit operations (insertions, deletions, substitutions, and shifts) that transform the output into a fluent and meaning-preserving translation. The minimum number of edit operations required to exactly match the reference translation, divided by the average number of reference words, then yields the HTER score. The average performance numbers of our system in terms of BLEU and HTER for various conditions are listed in Table 1; the performance is state-of-the-art and comparable to that of other systems on this task.

	NW	NG	BN	BC
Chinese				
BLEU	13.55	11.13	11.70	8.47
HTER	28.19	30.35	29.16	35.59
Arabic				
BLEU	22.88	10.03	16.06	12.41
HTER	17.70	33.56	28.92	38.23

Table 1: Average MT performance (BLEU(%) and HTER) of the system on various genres (NW = newswire, NG = newsgroups, BN = broadcast news, BC = broadcast conversations) and languages.

HTER scores were used as the performance scores in our error analysis. The number of available documents annotated for HTER, as well as the number of documents selected for this study, are shown in Table 2. Since time and resource constraints did not permit processing of all documents, a representative selection of the best/worst 20-50% of documents in a given category was used. For conditions with only a small number of available documents, near-complete coverage was sought.

	Arabic		**Chinese**	
	available	used	available	used
NW	44	28	35	17
NG	27	17	35	22
BN	20	8	16	16
BC	7	6	11	11

Table 2: Number of documents used for the error analysis.

Error Analysis Procedure

Our overall error analysis approach is as follows:

1. Define a list of potential factors influencing MT performance. This may include e.g. dialect, source, genre, ASR performance, etc.
2. For each segment in the MT input/output, extract quantitative or categorical measurements of each factor from the input document, either automatically, or aided by human annotation.
3. Measure the mutual information between each measurement and the MT performance score for that segment.
4. Rank factors in terms of highest-to-lowest mutual information.

Factors appearing at the top of the list can then be assumed to be more correlated with MT performance scores than factors at the bottom of the list.

Factors

The following features of input documents were established as potentially relevant to both audio and text documents:

1. **Genre**: one of the four genres mentioned above (BC, BN, NW, NG). Since current SMT systems are trained on large amounts of text data (and only small amounts of e.g. parallel transcriptions of conversations), it is likely that the genre will play a role in predicting MT performance.
2. **Source**: the identifier of the particular show or newspaper from which the document was extracted. This feature might reveal a bias of particular sources towards a vocabulary or style that is not handled well by the MT system. In the case of text documents this might indicate the preferences of individual authors; in the case of audio documents it could also indicate recording conditions or speaker effects.
3. **Target language model score**: the perplexity obtained by a well-trained target language model on the (manually created) reference translation of the document. Unigram, bigram and trigram scores are used separately.
4. **OOV rate**: the percentage of out-of-vocabulary (OOV) words in the input document, i.e. words that were not seen in the training data and thus are novel to the system.

5. **Style**: this feature indicates the style (spoken, written, mixed) of the document.
6. **Dialect**: the dialect of the source document.
7. **Names**: the percentage of correctly translated named entities in the translation output relative to the reference translation. Since many documents are from the news domain, where new names occur frequently, the failure to correctly translate named entities may be a significant contributor to poor overall MT performance.

Features 3 and 4 are intended to measure the mismatch between training and test data. A test document that receives high perplexity under a language model trained on a large training data set can be considered mismatched either in terms of topic/domain or in terms of style. Words from novel topics or domains that are not represented in the training data will cause the language model to backoff to the unknown word probability; similarly, word sequences caused by differences in style (e.g. spoken, conversational style as opposed to written text) will receive low probabilities.

A second way of measuring mismatch, in particular mismatch caused by topic/domain differences, is by computing the OOV rate. In languages with strong dialectal variation, the OOV rate may also indicate dialect effects.

For audio documents, we additionally extract:
8. **WER**: the word error rate of the ASR system for the particular MT segment. For Chinese, character error rate (CER) is used instead.
9. **% substitutions**: the percentage of substitutions in the ASR output.
10. **% deletions**: the percentage of deletions in the ASR output.
11. **% insertions**: the percentage of insertions in the ASR output.
12. **Dialect-ASR**: the dialect rating assigned to the ASR output (as opposed to the rating based on the reference transcription (feature 6 above)).
13. **Style-ASR**: the style rating for the ASR output
14. **Names-ASR**: the percentage of correctly translated names in the ASR output .

Features 1-4 measure the performance of the ASR front-end. We also include three binary features (**Δ Dialect**, **Δ Style**, **Δ Names**) indicating whether features 5, 6 and 7 differ from the corresponding ratings (features 12, 13, 14) based on the ASR reference transcription (e.g. whether the style of the ASR output was judged differently from that of the ASR reference transcription). If ratings differ, they may indicate problems with the ASR component.

Measurements of all factors, as well as HTER scores, are computed at the document level. Although it would be desirable to choose finer-grained segments (sentences or even phrases), a document-level segmentation was the only segmentation observed by both the ASR and the MT components, as well as the HTER annotation. Internally, all components use different sub-document segmentations, which precluded the use of smaller segments. It should be noted that MT performance can vary within a single document; however, our analysis will only consider the average performance over the entire document.

Measurements

Most of the factors listed above can be measured automatically. For our particular task, genre and source information were available from the information distributed with the test documents. Source and target language model scores were supplied in the form of separate unigram, bigram and trigram perplexity scores obtained by large-scale language models trained on billions of words of training data (primarily newstext but also including conversational data). The target language model scores were provided by the same English language model for both Arabic and Chinese. The source language models were the language models used by the respective Arabic and Chinese ASR front-ends in the system. Text preprocessing and tokenization were applied to the target/source texts to match the preprocessing required by the language models. The OOV rate was obtained from the best individual SMT system in the system combination; this was the system developed by RWTH Aachen. The ASR performance scores were computed by standard scoring of the ASR hypotheses against the reference transcriptions of the audio files.

Dialect and style ratings, as well as the percentage of correctly translated names, were determined from human annotations specifically performed for this error analysis study. For Chinese, four dialectal categories were established: Mainland Chinese, Hong Kong Chinese, Taiwanese Chinese, and “neutral/can't tell”. Style was categorized as “spoken”, “written", or “mixed”.

For Arabic, dialect was annotated per word, i.e. each word in the document was assigned a degree of “dialectalness” ranging between 0 and 3. Level 0 is the default assigned to all pure Modern Standard Arabic (MSA) words and dialectal words that are MSA-like (these are words that are historically MSA and have remained phonologically unchanged or have slight phonological changes that are not seen in the written form). Level 1 of dialectness is given to words that have non-standard spelling: this category includes both spelling errors and dialectal words which are recognizable cognates of MSA words, e.g. هدا (instead of هذا ‘this’). Level 2 is given to words that could be Level 0 except for the presence of one of a special set of dialectal affixes that do not exist in MSA. For example, the +ب present tense prefix in Levantine and Egyptian Arabic which could attach to the otherwise Level 0 word يكتب yielding its dialectal variant: بيكتب. Finally level 3 is assigned to words that are completely dialectal regardless of what kind of morphology they exhibit.

Additionally, segment-level judgments of dialectalness ranging between 0 (perfect MSA) and 4 (pure dialect) were assigned. The difference between these two types of annotations is that segments can be judged dialectal even though all of its component words taken in isolation would be judged as MSA -- this may include e.g. certain idioms or colloquialisms that would not occur in a pure MSA text. An average dialectalness score for the entire document can then be computed either from the segment-level or the word-level ratings. For the present analysis the average word-level score was used. Style was not annotated separately for Arabic since it was felt that it coincided fairly closely with dialectalness (spoken documents being automatically more dialectal).

Arabic-English name annotation involved identifying names in the Arabic sentences and matching them with their translation in the target English sentences. We can distinguish seven different cases:

0. The word is not a name (default).
1. The word is not a name but is erroneously translated as a name in English.
2. A name that is not translated at all.
3. A name that is not translated as a name.
4. A name that is translated as a different (incorrect) name.
5. A name and it is translated into English correctly EXCEPT that it is not capitalized.
6. A name that is translated correctly and capitalized correctly.

Of these, 1-5 constitute translation errors. The sum of all translation errors over the entire document, divided by the total number of names, is the name error rate used in the error analysis.

For Chinese, a slightly different name annotation scheme is used which rewards partially correct name translations. Due to the character-based Chinese script, names consisting of multiple characters may have translations where one character was correctly translated but the others were not. A “correct” score is assigned to each correctly translated part.

Human Annotation

The human annotations for Chinese were performed by 6 native Chinese speakers (graduate students and postdocs) at the University of Washington. Each annotator processed a different subset of documents. Due to time and resource constraints, individual documents were not annotated multiple times, and inter-annotator agreement thus was not measured. However, annotations were spot-checked by other native speakers to ensure the correctness of the annotations. Each document took approximately 20 minutes on average to be annotated, but annotators noted a large variance in the amount of time required. Annotation speed mainly depended on whether the document was an audio or text document (audio being harder), document length, and on the font encoding (traditional vs. simplified Chinese font).

The Arabic annotations were carried out at Columbia University by four native speakers of Arabic. Name annotation was divided among all four. Dialectness annotation was completed by one annotator only who received special training. The dialect annotation took 23 minutes on average per document. Documents that were mostly in MSA were the easiest and fastest to finish, while documents with many dialectal words took longer. The basic assumption in the annotation of dialects is that a word is in MSA unless it exhibits a special feature as discussed above.
Name annotation took more time on average (~ 67 minutes per documents), This is due to the complexity of the task which include recognizing the names in the source and verifying their presence in the target machine translation (as opposed to the dialectness identification task which is a monolingual task). Similarly to dialectness annotation, some documents took much longer (upwards of two hours per document) whereas others where much faster to annotate. Speech recognition errors added to machine translation errors made this task especially hard.

For both languages and all annotation types, individual documents were not annotated multiple times and inter-annotator agreement thus was not measured due to time and resource constraints. However, annotations were spot-checked by other native speakers to ensure the correctness of the annotations.

Mutual Information

In order to measure statistical dependencies between the measurements described above and HTER we chose mutual information (MI). The (discrete) mutual information between two random variables *X* and *Y* is defined as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

It is the most general way of measuring dependencies between two random variables since it can capture both linear and non-linear dependencies. Mutual information expresses the degree to which knowledge of one variable reduces the uncertainty about the other. Alternatively, it can be thought of as the distance (KL divergence) between the joint distribution *p(X,Y)* and the product of their marginal distributions, *p(X)p(Y)*, the implication being that the distance should be 0 if *X* and *Y* are entirely independent.

Since our measurements involve both categorical and continuous values, we use discrete mutual information and perform histogram-based binning of the continuous values prior to computing the mutual information. Binning is based on the Freedman-Diaconis rule (Freedman & Diaconis 1981), which calculates the bin width *w* as:

$$w = 2 * IQR(x) N^{-\frac{1}{3}}$$

where *N* is the number of samples in the population *x* and IQR is the interquartile range (the range between the first and third quantiles) of *x*.
After the mutual information has been computed between each factor measurement and HTER, factors are ranked from highest to lowest mutual information. The factors appearing at the top of the list thus are the factors most predictive of HTER. At this point, we do not yet make use of any statistical significance tests to determine the significance of different ranks.

System Analysis

In this section we present the application of our error analysis framework to our specific MT system. We

computed the mutual information based ranking of input document features in three different ways:
(a) across all documents pertaining to a given language (Chinese or Arabic),
(b) for the set of audio vs. text documents for each language (each set comprising two genres), and
(c) separately for each genre and each language.

This procedure yields a successively finer-grained analysis; however, the number of documents (and thereby the sample size from which the mutual information is computed) becomes smaller as well. In each case, binning of continuous values is redone based on the changed sample.

Chinese

The mutual information analysis for the set of all Chinese documents is performed only on those features that are defined for both text and audio documents, i.e. excluding e.g., the measurements of ASR performance. Based on these common features, the ranking of different factors is as shown in Table 3.
According to this analysis, source and genre are the most important factors, followed by unknown words (as indicated by the target unigram score, OOV rate and the percentage of correctly translated names). Style and dialect, by contrast, do not seem to play a significant role.

Rank	Factor
1	Source
2	Genre
3	Target unigram score
4	OOV rate
5	Names
6	Target bigram score
7	Target trigram score
8	Style
9	Dialect

Table 3: Ranking of factors for Chinese documents (cross-genre).

It is not surprising that genre emerges as one of the strongest predictive factors: since all component systems contributing to the combined system under investigation have mostly been trained on text data (newswire text, and parliamentary proceedings). Therefore, they generally produce low-quality translations when presented with unstructured text sources such as newsgroups data, or with ASR output. The source effect could be due to specific styles or vocabularies employed by different news sources, as well as a speaker or an acoustic effect in the case of audio files. Due to the limited sample size, the source effect could also be an individual document effect. Further investigations on a larger data set will be required to resolve this question.

Text and audio genres were subsequently analyzed separately. Table 4 shows the ten top-ranking factors for each condition (note that text documents only have nine factors in total). We see that for text translation the percentage of correctly translated names plays a significant role, as well as the source and the presence of unknown words. For audio documents, source and unknown words seem to be relevant, in addition to errors (substitutions) in the ASR output.

Finally, each genre was analyzed separately. The results are shown in Tables 5 and 6. For text genres (Table 5), name translation again emerges as one of the most important factors, as well unknown words in the test data.

Rank	Text	Audio
1	Names	Source
2	Source	Target unigram score
3	OOV rate	Source unigram score
4	Target unigram score	% substitutions
5	Target trigram score	CER
6	Target bigram score	Target bigram score
7	Genre	Source bigram score
8	Dialect	Dialect
9	Style	% insertions
10	---	Target trigram score

Table 4: Top-ranking factors for Chinese audio and text documents.

.
This is not a surprising result because names are very frequent in newstexts, and their mistranslation is a significant source of errors. In addition, there seems to be a strong source effect for Chinese newsgroups documents.

Rank	Newswire	Newsgroups
1	OOV rate	Source
2	Names	Names
3	Target unigram score	Target unigram score
4	Target bigram score	Target bigram score
5	Target trigram score	Target trigram score
6	Dialect	Dialect
7	Source	OOV rate
8	Style	Style

Table 5: Ranking of factors for Chinese text documents.

Rank	Broadcast News	Broadcast Conversations
1	Source bigram score	Source
2	Target unigram score	Style
3	CER	% substitutions
4	% deletions	CER
5	Source unigram score	Names-ASR
6	% substitutions	Δ Dialect
7	Source trigram score	% deletions
8	Source	% insertions
9	Style of source	Names
10	Dialect of source	Dialect-ASR

Table 6: Ten top-ranking factors for Chinese audio documents.

For audio genres (Table 6) we also observe source effects for broadcast conversations, which could be acoustic in nature in this case (i.e. sensitivity to specific speakers or acoustic recording conditions). Furthermore, ASR performance is a strong predictor of MT performance for both broadcast news and broadcast conversations, with deletions being more dominant than other ASR errors in the case of broadcast news. Training/test data mismatch (as indicated by source and target language model scores) is also relevant in the broadcast news condition. It is interesting to note that names are not among the top ten factors for broadcast news, though they do show up in the list for broadcast conversations.

Arabic

We conducted analogous analyses for Arabic. The cross-genre ranking is shown in Table 7. Overall, source is again one of the dominant factors, along with the OOV rate and correct name translation. Contrary to our initial expectations, dialect did not seem to be a major factor although Arabic is known to have much dialectal variation.

Rank	Factor
1	Source
2	OOV rate
3	Target trigram score
4	Names
5	Genre
6	Target unigram score
7	Target bigram score
8	Dialect

Table 7: Cross-genre factor ranking for Arabic.

A comparison of the factor ranking for text vs. audio ranking is shown in Table 8. In both conditions we find strong source and OOV effects. In addition to name translation in the text condition we find a significant effect from deletions in the ASR output for the audio condition (much stronger than the effect from either substitutions or insertions). Dialect as a factor influencing MT performance does appear in the top ten list in both cases but, again, has a lower rank than expected. The use of segment-level as opposed to word-level dialect scores (see Section "Measurements" above) did not yield a different ranking.

The analysis of the separate genres (Tables 9 and 10) provides a more detailed picture: In the text genres, source is an important feature for both newswire and newsgroups data; so is the percentage of correctly translated names.
OOV words seem to be much more prevalent in the newsgroups data, which is plausible considering that unstructured texts such as emails and blogs contain many neologisms and non-standard words. When looking at the audio genres, we find that the rate of deletions in the ASR output is (quite surprisingly) the most important factor for broadcast news documents, but it is much less important for broadcast conversations, which are dominated by source and OOV effects.

Rank	Text	Audio
1	Source	Source
2	OOV rate	OOV rate
3	Names	% deletions
4	Target bigram score	Source trigram score
5	Target trigram score	Source bigram score
6	Genre	WER
7	Target unigram score	% substitutions
8	Dialect	Dialect
9	----	% insertions
10	----	Target bigram score

Table 8: Factor rankings for Arabic text and audio documents.

Rank	Newswire	Newsgroups
1	Source	Source
2	Names	OOV rate
3	Target unigram score	Names
4	Target trigram score	Target bigram score
5	Dialect	Target unigram score
6	Target bigram score	Target trigram score
7	OOV rate	Dialect

Table 9: Factor rankings for Arabic text documents.

A closer analysis of this phenomenon revealed that the rate of deletions for broadcast news documents was highly variable (more variable than the rate of insertions or substitutions) and was extremely high for certain shows and speakers. It turned out that those speakers had extremely low-amplitude signals, which were either due to inherently quiet voices or recording conditions. This problem was not handled by the adaptation component in the ASR system. The rate of deletions in the Chinese system was also high; a similar analysis revealed problems not with adaptation but with acoustic segmentation. Both of these problems have since been addressed by employing more sophisticated acoustic processing models. This is thus a concrete example of a system weakness in the ASR-MT pipeline which has been revealed by our error analysis procedure.

Rank	Broadcast News	Broadcast Conversations
1	% deletions	Source
2	% substitutions	OOV rate
3	WER	Target trigram score
4	Names	Target bigram score
5	Target unigram score	% substitutions
6	Source	Target unigram score
7	OOV rate	WER
8	Source trigram score	% deletions
9	Source bigram score	Dialect
10	Dialect	Source trigram score

Table 10: Ranking of the top ten factors for Arabic audio documents.

Rank	Broadcast Conversations
1	Source
2	OOV rate
3	Target trigram score
4	Target bigram score
5	% substitutions
6	Dialect
7	Target unigram score
8	WER
9	% deletions
10	Source trigram score

Table 11: Ranking of the top ten factors for broadcast conversation documents with segment-level dialect scores.

We also ran the analysis on the individual genres with segment-level as opposed to word-level dialect scores and found that in broadcast conversations, the two different annotation schemes do make a difference: the ranking of dialect as a factor affecting output performance changes from position 9 to position 6, indicating that a segment-level annotation may be more useful in capturing dialect effects in conversations. Whereas a word-level annotation only considers the percentage of dialectal words in a text, the segment-level annotation is more apt to capture the speaker's intent.

Conclusions

We have presented a general framework for semi-automatically analyzing characteristics of input documents to MT systems that determine output performance. The framework was illustrated here for a specific set of languages, genres, and input text features. However, the method is more general and can be applied to other languages pairs and system measurements. It has the potential of being fully automated by replacing manual annotations with automatic annotations in the future, e.g. named entity recognition and classification, automatic dialect classification, or automatic genre detection. Furthermore, it can be refined by including additional factors such as speaker identity, direct acoustic measurements extracted from the speech signal (e.g. SNR or speaking rate), more fine-grained representations of different sources (including e.g. author ids) or separate scores from POS-based vs. word-based language models, in order to separate out style vs. domain effects. Another issue to be improved is the type of document segmentation used. Whereas we have used document-level segmentations in this study, segmentation at a more fine-grained level would yield a more detailed picture of input-output effects and would provide a larger sample for computing mutual information values. However, we have seen that even for a small sample size, the analysis performed here provides useful guidance as to which problems should be investigated in more detail -- the problems of ASR deletions, for instance, would have been difficult to diagnose just by manually analyzing the MT output without any prior guidance. Thus, our method is useful for quickly identifying which components need to be looked at more carefully and which problems need to be followed up on with more detailed manual analysis.

Acknowledgements

This work was funded by DARPA Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these agencies. We thank Evgeny Matusov for providing part of the input data and Karim Filali for help with the mutual information software.

Bibliographical References

S. Banerjee and A. 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*, Ann Arbor, MI.

G. Doddington (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *Proceedings of the ARPA Workshop on Human Language Technology (HLT)*, San Diego, CA, USA, pp. 128-132.

D. Freedman and P. Diaconis (1981) On the histogram as a density estimator: L_2 theory. *Probability Theory and Related Fields 57(4)*, 453-476.

Philip Koehn (2003) *Noun phrase translation*. PhD Thesis, University of Southern California.

E. Matusov N. Ueffing and H. Ney (2006) Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proceedings of EACL 2006*, pp. 33-40, Trento, Italy, April 2006.

S. Niessen, F.J. Och, G. Leusch, and H. Ney (2000) An evaluation tool for machine translation: fast evaluation for MT research. In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pp. 39-45.

F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin and D. Radev (2003) *Syntax for Statistical Machine Translation*. Final Report of the Johns Hopkins 2003 Summer Workshop.

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

M. Popovic and H. Ney (2006) Error Analysis of Verb Inflections in Spanish Translation Output. In: *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, pp. 99-103.

M. Popovic, A. de Gispert, D. Gupta, P. Lambert, H. Ney, J.B. Mariño, M. Federico, and R. Banchs (2006) Morphosyntactic Information for Automatic Error Analysis of Statistical Machine Translation Output, *Proceedings of the ACL Workshop on Statistical Machine Translation*, pp. 1-6.

M. Snover, B. Dorr, R. Schwartz, J. Makhoul, L. Micciulla and R. Weischedel (2005) *A study of translation error rate with targeted human annotation*, University of Maryland Technical Report, UMIACS-TR-2005-58.