Statistical Machine Translation for Biomedical Text: Are We There Yet?

Cuijun Wu, MA[1], Fei Xia, PhD[1], Louise Deleger, PhD[2], Imre Solti, MD, PhD, MA[2*]
[1]University of Washington, Seattle, WA; [2]Cincinnati Children's Hospital Medical Center, Cincinnati, OH

{cuijunwu,fxia}@uw.edu; {louise.deleger,imre.solti}@cchmc.org
* Corresponding Author

Abstract

In our paper we addressed the research question: "Has machine translation achieved sufficiently high quality to translate PubMed titles for patients?". We analyzed statistical machine translation output for six foreign language - English translation pairs (bi-directionally). We built a high performing in-house system and evaluated its output for each translation pair on large scale both with automated BLEU scores and human judgment. In addition to the in-house system, we also evaluated Google Translate's performance specifically within the biomedical domain. We report high performance for German, French and Spanish -- English bi-directional translation pairs for both Google Translate and our system.

Introduction

One of the aims of patient centered medicine is to empower the patients in the medical decision making. According to the US Census Bureau, 18 percent (47 million people) of the US population aged five and over reported they spoke a language other than English at home in 2000[1]. To fulfill the promise of patient centered medicine for non-English speaking patients in the United States (US) it is immensely valuable to make English language biomedical text available in foreign languages. The Census Bureau estimates that there are about 45 million Hispanics living in the United States and many of them are Spanish-only speakers[2]. Not only Spanish native speakers, but other US residents would also benefit from accessing biomedical information in their native tongue even if they can communicate in English.

In addition to patient centered medicine, clinical trials require the translation of biomedical text, as well. An increasing number of clinical trials require cross-border and cross-language enrollment in order to have a sufficiently diverse representation of the human gene pool. There is also a growing need to collect and aggregate disease-specific information across countries and continents to achieve meaningful sample size for rare diseases. In case of international research, much of the clinical information is locked into free text in different languages. Accessing this information, either automatically by Natural Language Processing tools or by human investigators is much easier if automated, timely, high quality and scalable translations are available.

In the past two decades, statistical machine translation (SMT) has become the dominating approach to machine translation (MT) due to its robustness, good performance, and the fact that it does not require manually crafted rules[3]. There are state of the art translation engines that were developed for general translation purposes. One of the most sophisticated publicly accessible machine translation engine is Google's Google Translate[4]. It is unclear if the Google Translate system has any specific training for the biomedical domain. To our knowledge, our work is the first evaluation of Google Translate for the biomedical domain

In this paper we present the results of our experiments to evaluate a state of the art general-purpose (Google Translate) and an in-house developed biomedical field focused statistical machine translation system. In our work we build on the success of publicly released statistical machine translation components and downloadable parallel biomedical corpus. We evaluate the performances of Google and our system against the human generated parallel corpora using an automated scoring system. To round out the evaluation process we employed human annotators to judge the quality of the machine translation system's output.

In the "Background" section we will describe the most relevant machine translation efforts in the non-biomedical domain and some of the earlier translation works that are focused on biomedical text. In "Methods" we will provide a detailed description of the task, the data and evaluation approaches. In "Results" we will show the findings from the automated and human translation evaluations. In "Discussion" we will analyze the results and finally we will present the "Conclusions".

Background

There are two streams of related work that we intend to cover for this paper. First, we will describe a few selected general-purpose machine translation works that are most relevant for our topic and evaluation approaches. Second, we will provide details of the two biomedical focused translation efforts known to us.

For MT evaluation, there are two types of evaluation: human evaluation and automatic evaluation. In human evaluation, bilingual speakers are presented with source sentences and translations produced by an MT system, and asked to judge the fluency and adequacy of the translations in a 1-5 scale[5]. This approach is intuitive and the results are easy to understand. However, human evaluation is slow, labor intensive, expensive, and cannot be reused. It is also subjective and may not be sensitive to small changes of MT quality. Because of these disadvantages, human evaluation cannot be used to monitor the effect of daily changes to an MT system in order to weed out bad ideas from good ideas[6].

To address these limitations, Papineni and his colleagues proposed an automatic measure called BLEU[6]. The main idea behind the measure is that the closer a MT translation is to a professional human translation, the better it is. To calculate the BLEU score, MT translations are compared with reference human translations and n-gram (n=1,2,3,4) precisions are calculated, where n-gram precision is the percentage of word n-grams in an MT translation that also occur in the corresponding human reference translations. BLEU score is defined to be the geometric mean of n-gram precisions multiplied by the brevity penalty (which is used to penalize an MT translation that is shorter than the reference translations).

Very often the human reference translations are already available from various bilingual text. Compared to human evaluation, calculating BLEU is cheap and quick, and it can be done frequently to assess daily changes of MT systems. BLEU also correlates well with human judgment. Recently, other automatic measures such as TER and METEOR have been proposed[7, 8]. In this paper, because BLEU is the most well-known and commonly used automatic measure in the SMT community, we will use it for evaluation, in addition to human judgment.

We know of only one translation tool, or more accurately a cross-language tool that was developed specifically for the PubMed text corpus. BabelMeSH was developed for Medline/PubMed[9]. BabelMeSH is a cross-language tool for searching Medline/PubMed articles in the user's native language. However, the tool is not intended as a full-text translation engine. It focuses only on searching Medical Subject Header (MeSH) terms in PubMed by utilizing the foreign language entries in the Unified Medical Language (UMLS) Metathesaurus.

Turner et al, are working on a public health documentation focused machine translation system[10]. Their goal is to improve the availability of health materials for individuals with Limited English Proficiency, and develop fundamentally new machine translation technology designed to adapt generic systems to the health care domain. Ultimately, they want to eliminate health disparities caused by language barriers and improve access to pertinent multilingual health information for patients.

In our paper we address the research question: "Has machine translation achieved sufficiently high quality to translate PubMed titles for patients?". In more general terms, we will answer the question: "Are we there yet?". Is it possible to start using statistical machine translation systems to generate high quality, large scale PubMed title translations? We will evaluate the output of Google Translate and our in-house built system that we will call from now on as ***BioMT***. When we started this research we were specifically interested in the human judged quality of system generated translations.

Methods

Data

We selected the freely leasable database of Medline/PubMed articles as the foundation of our corpora. The biomedical parallel corpus was constructed from the foreign language titles and their corresponding English translated titles of Medline/PubMed articles. The database had (as of March 2010) over 17 million titles and covered 55 languages with the vast majority being English-only titles.

For the parallel corpora we extracted French, Spanish, German, Hungarian, Turkish and Polish titles and corresponding human English translations. The Medline/PubMed database consists of XML files that include XML tags for foreign titles (so-called Vernacular Titles) and the English translations (so-called Article Titles). Figure 1 presents an example for a German-English title pair.

<VernacularTitle>Pathologisches Kaufen und psychische Komorbidität.</VernacularTitle>

<ArticleTitle>[Compulsive buying and psychiatric comorbidity]</ArticleTitle>

Figure 1. An example for vernacular (German) and corresponding English title

For the pre-processing steps, we used regular expressions to find the <ArticleTitle>, and corresponding <VernacularTitle> tags and extract the contents from the XML files. We randomly selected 80% training, 10% test, and 10% development data. Table 1 shows the descriptive statistics of our parallel training, development and test corpora (after filtering out training sentences that were longer than 40 words).

Table 1. Number of human translated (reference) titles per studied languages

	French	Hungarian	Polish	Spanish	German	Turkish
Training	443,862	26,345	121,829	187,059	565,006	5,329
Test	55,598	3,296	15,249	23,411	70,626	668
Development	55,598	3,296	15,249	23,411	70,626	668

Building the In-House Machine Translation System

To build BioMT we used the Moses toolkit, a state-of-art open-source phrase-based SMT system and followed its step by step guide[11] Moses builds a translator for S=>T (S is the source and T is the target language) in three stages: training, tuning and testing.

At the training stage, Moses first learns word-to-word translation and distortion models from the training data using IBM Models 1-5, then uses the models to find word alignment between each sentence pair in the training data, and next uses the word alignment to build a phrase table and reordering model[12]. The phrase table stores the probability that a source phrase translates into a target phrase. "Phrase" in this context refers to a word n-gram, not necessarily a linguistic phrase. A reordering model captures how likely the source phrases are reordered in the target side. Finally, Moses uses the SRLIM package[13] to build an n-gram language model from the target side of the training corpus.

A tri-gram language model was trained on the target side of the training parallel corpus using the SRILM package. The translation and re-ordering model relied on "intersect" symmetrized word-to-word alignments. That is, each word alignment can be seen as a set of source word - target word pairs. Moses takes the intersection of the two sets, and uses that as the final word alignment.

The goal of the tuning stage is to learn good weights of the translation, reordering and language models. The tuning is done by running the machine translation system with various weight combinations on a set of new sentences and choosing the combination that produces good translation results (the evaluation is described below in more detail). For tuning we experimented with several tuning sizes: we used the first 200/300/400/500 lines of the development data set.

The last stage is testing (also called decoding). Moses uses the models learned in the training stage and model weights chosen in the tuning stage to translate sentences in the test data. We used the default settings recommended by Moses for the decoder.

Collecting Output for the Google Translator

To compare the performances of our system and the Google translation engine we submitted the test corpora (Table 1) for each language pairs and each translation direction (that is, Foreign to English and English to Foreign) to Google Translate. We used the publicly available Google Translator API to connect to the Google service[14].

Testing the Effect of Training Corpus Size

To measure the impact of the size of the training corpus on the performance of the machine translation system, we experimented with different training sizes by changing the number of foreign titles with corresponding human reference translations in the training corpus. We measured the BLEU scores while evaluating on the same test corpus.

Evaluation Methods

We implemented both automated (BLEU score) and human evaluation processes to measure translation quality. While BLEU is a standard method of evaluation in the general field of machine translation, we are aware of the importance of generating linguistically and culturally appropriate translations for patients[15]. In order to measure the cultural and linguistic appropriateness of the translations we hired bilingual and monolingual judges. We had two bilingual judges for Spanish, Hungarian and Polish each, and one bilingual judge for French. We had no bilingual judges for German and Turkish so English to German and English to Turkish translations were not evaluated. We also hired two monolingual (English-only) judges who evaluated Foreign to English translations across all the six languages.

The bilingual judges evaluated the translation quality in both directions of the translations (Foreign to English and English to Foreign). The monolingual judges evaluated the quality only of the Foreign to English translations. All judges were untrained in translation evaluation. The monolingual judges were recent BSc (Statistics and Anthropology) graduates. The bilingual judges were all native speakers of the evaluated language and lived in the US. The bilingual judges had Masters or Doctoral degree as their highest educational diploma. One of the Hungarian and the single French judge are co-authors of the paper but none of the other judges had any involvement in the study other than evaluating the quality of the translations.

For human evaluation, 100 titles were randomly selected from the test set for each language with their corresponding foreign and English reference translations. The corresponding 100 BioMT and Google translations were selected as well. The source titles, the reference (human) translations and the two systems' (BioMT and Google) outputs were presented to the judges who scored the translation quality for "Fluency" and "Content" on a 1-5 scale (1/worst and 5/best). The judges were also asked to indicate which translation they considered better. Figure 2 demonstrates an example from the French title and translation set with corresponding questions and scores from the judge.

9_8 SOURCE: La sociothérapie: une nouvelle thérapie?
9_8 HUMAN: Sociotherapy: a new therapy?
9_8 SYS1: social therapy : a new therapy ?
9_8 Fluency (1(worst) - 5(best)): 5
9_8 Content (1(worst) - 5(best)): 4
9_8 SYS2: sociotherapy : a new therapy ?
9_8 Fluency (1(worst) - 5(best)): 5
9_8 Content (1(worst) - 5(best)): 5
9_8 Which is better? Sys1 vs Sys2 (1 vs 2): 2
Figure 2. Snippet from the French to English scoring file

"9_8" indicates the title number in the 100-title set and the example illustrates that the judge gave 5 for "Fluency" for both system's output and scored the "Content" 4 and 5 while indicated that the second system provided a better translation. The order of printing BioMT's and Google's outputs were randomly switched for each of the 100 titles to avoid developing a bias against either "SYS1" or "SYS2" as presented in the files. While the investigators kept track which system corresponded to Google and BioMT the judges were unaware of this information. The scores were collected with an automated process.

The judges were instructed to evaluate the translation characteristics as follows: "Content: How well the main message of the source sentence is communicated in the translation even if the translation's fluency is terrible." and "Fluency: How human like is the translation as a sentence in the target language?". To answer the last question, "Which is better? Sys1 vs Sys2 (1 vs 2)):" the judges could answer 1 (SYS1 is considered a better translation), 2 (SYS2 is considered a better translation) or 0 (both translations are considered the same quality). Scores of "0" were discarded before running a Chi-square analysis on the scores for the fifth question.

Results

The following legend applies to each table and figure with language pairs indicated (FtE = French to English, EtF = English to French, HtE = Hungarian to English, EtH = English to Hungarian, PtE = Polish to English, EtP = English to Polish, StE = Spanish to English, EtS = English to Spanish, GtE = German to English, EtG= English to German, TtE = Turkish to English, EtT= English to Turkish).

Automated Evaluation:

Table 2 shows the BLEU scores for each pair of translations. To produce the BLEU results shown in Table 2, the BioMT system was trained on the maximum number of available training corpora. For example, in order for BioMT to generate 45.46 BLEU score for the French to English translation direction (as measured on 55,598 title translations in the French test corpus), the BioMT system was trained on all 443,862 training titles and their corresponding human (reference) translations.

Table 2. BLEU Score obtained by the two systems for each language translation

	FtE	**EtF**	**HtE**	**EtH**	**PtE**	**EtP**	**StE**	**EtS**	**GtE**	**EtG**	**TtE**	**EtT**
Google	37.74	34.95	19.08	8.08	29.98	17.54	45.65	44.14	36.39	23.2	26.52	13.63
BioMT	45.46	46.54	17.35	10.88	36.04	31.7	47.64	49.32	39.63	34.48	17.33	15.4

Figure 3 shows the BLEU scores for each language pair for both the Google (G) and the BioMT systems.

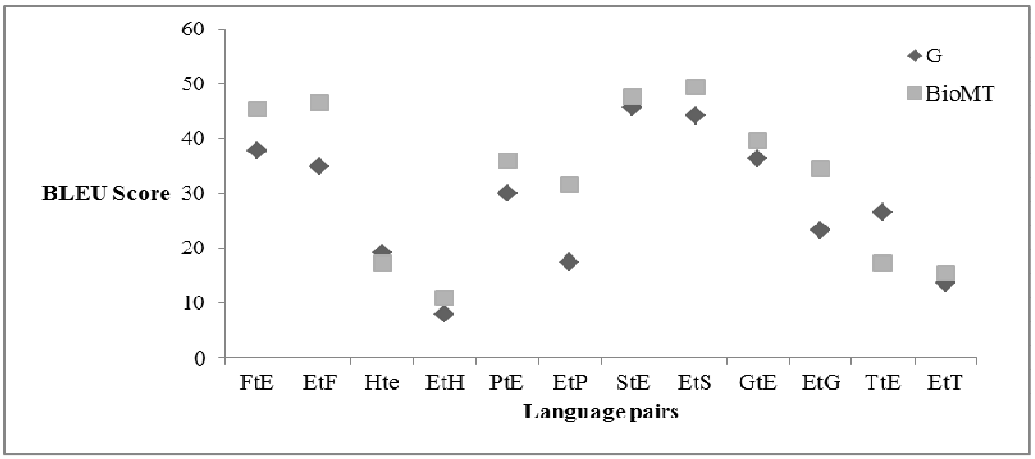

Figure 3. BLEU scores per languages and systems

Table 3 shows the impact of the training corpus' size on the BLEU evaluation scores for BioMT.

Table 3. BLEU score achieved by BioMT trained on various corpus sizes

	FtE	EtF	PtE	EtP	StE	EtS	GtE	EtG
50,000	31.34	28.14	29.09	23.01	36.43	36.28	23.62	16.40
100,000	36.55	34.62	34.85	29.29	44.47	44.83	28.82	20.21
150,000	38.75	38.41			47.72	49.01	30.89	23.03
200,000	41.29	40.20					32.82	25.45
250,000	42.38	42.18					33.58	27.31
300,000	43.68	44.45					34.28	28.59
350,000	44.23	45.81					35.03	29.42
400,000	44.95	46.64					36.46	31.03
450,000							36.83	32.16
500,000							38.59	33.15
550,000							38.96	33.67
All data	45.76	47.24	36.04	31.44	48.63	50.84	39.07	34.06

Figures 4 and 5 visualize the impact of the training corpora on the automated BLEU evaluation scores.

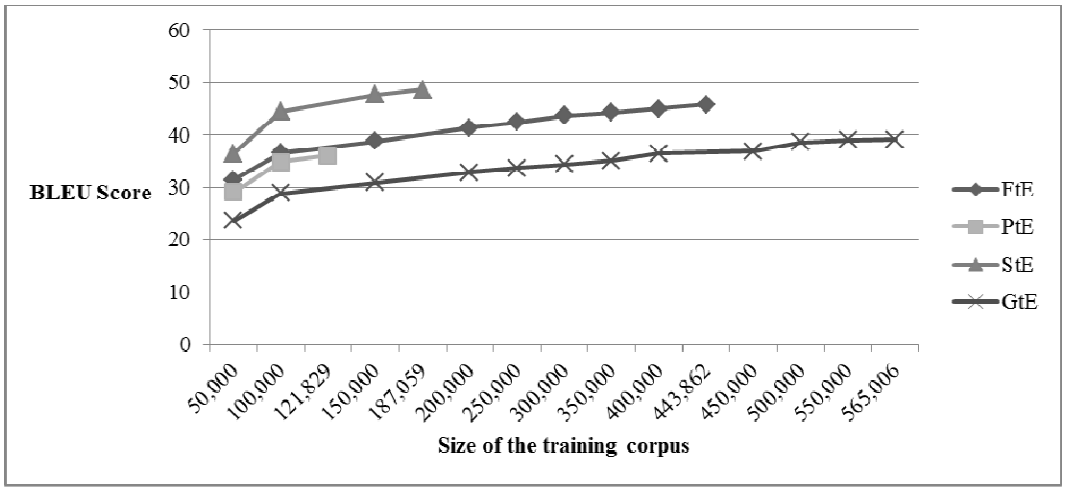

Figure 4. BLEU scores for Foreign to English translations (BioMT system)

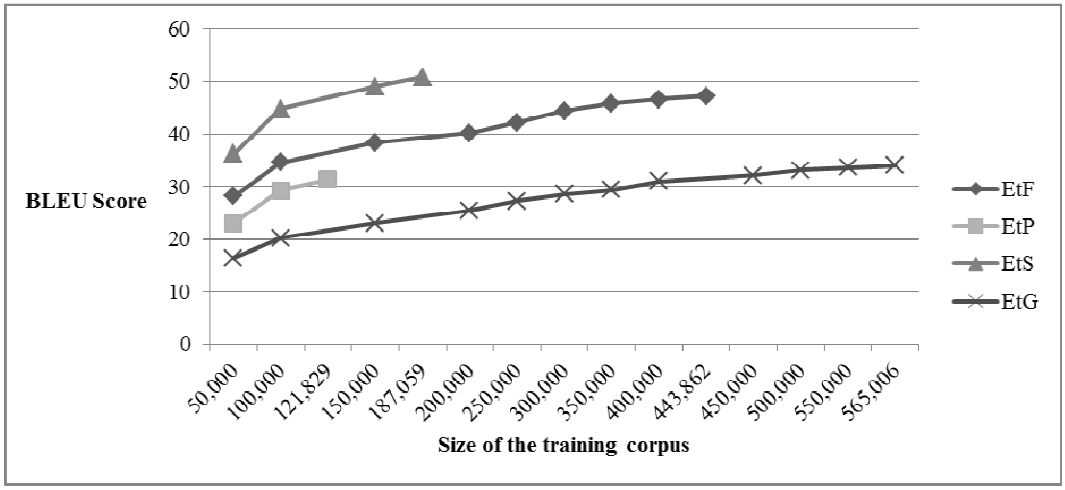

Figure 5. BLEU scores for English to Foreign translations (BioMT system)

Human Evaluation:

Including the bidirectional and monolingual scoring we collected 26 text files from the judges. Each file included the original text and the judges' scores (as presented in Figure 2 above). The judges made five scoring decisions for each of the 100 titles and corresponding translations for each of the files. Altogether the human judges made 13,000 scoring decisions (26*100*5). Table 4 shows the number of scoring decisions for each direction of the studied translations. The number of scoring decisions per language pair depends on the availability of bilingual judges as described in the Methods section.

Table 4. Number of scoring decisions for each type of language translation

	FtE	**EtF**	**HtE**	**EtH**	**PtE**	**EtP**	**StE**	**EtS**	**GtE**	**TtE**
Number of scoring decisions	1,500	500	1,500	1,500	1,500	1,500	1,500	1,500	1,000	1,000

Table 5 presents the averages of human judgment scores (both mono and bilingual when it was available) for the fluency and content of each translation per machine translation system. The table also presents the 95 percent confidence interval boundaries for the means. Boldface font type in the "Mean" column indicate a statistically significant difference (tested by non-overlapping 95 percent confidence intervals) in favor of the particular system.

Table 5. Human judgment for the two systems for each translation

		Google				BioMT			
		Mean	Std Error	95% confidence interval		Mean	Std Error	95% confidence interval	
				Lower	Upper			Lower	Upper
FtE	Fluency	4.133	0.064	4.008	4.259	4.163	0.064	4.038	4.289
	Content	4.107	0.064	3.981	4.232	4.123	0.064	3.998	4.249
EtF	Fluency	3.970	0.111	3.752	4.188	4.020	0.111	3.802	4.238
	Content	4.170	0.111	3.952	4.388	4.120	0.111	3.902	4.338
HtE	Fluency	**2.748**	0.056	2.639	2.856	2.065	0.056	1.956	2.174
	Content	**2.668**	0.056	2.559	2.776	2.020	0.056	1.911	2.129
EtH	Fluency	2.191	0.079	2.037	2.345	2.015	0.079	1.861	2.169
	Content	2.196	0.079	2.042	2.350	2.000	0.079	1.846	2.154
PtE	Fluency	3.838	0.056	3.729	3.946	3.885	0.056	3.776	3.994
	Content	3.673	0.056	3.564	3.781	3.625	0.056	3.516	3.734
EtP	Fluency	3.410	0.079	3.256	3.564	3.600	0.079	3.446	3.754
	Content	3.370	0.079	3.216	3.524	3.430	0.079	3.276	3.584
StE	Fluency	4.710	0.055	4.603	4.817	4.542	0.055	4.435	4.650
	Content	4.587	0.055	4.480	4.695	4.397	0.055	4.290	4.505
EtS	Fluency	4.645	0.077	4.493	4.797	4.605	0.077	4.453	4.757
	Content	4.675	0.077	4.523	4.827	4.640	0.077	4.488	4.792
GtE	Fluency	4.300	0.079	4.146	4.454	4.215	0.079	4.061	4.369
	Content	4.005	0.079	3.851	4.159	3.850	0.079	3.696	4.004
TtE	Fluency	**3.180**	0.077	3.028	3.332	2.270	0.077	2.118	2.422
	Content	**3.575**	0.077	3.423	3.727	2.435	0.077	2.283	2.587

Figure 6 presents the mean fluency and content scores. The results are plotted per translation pairs for both systems.

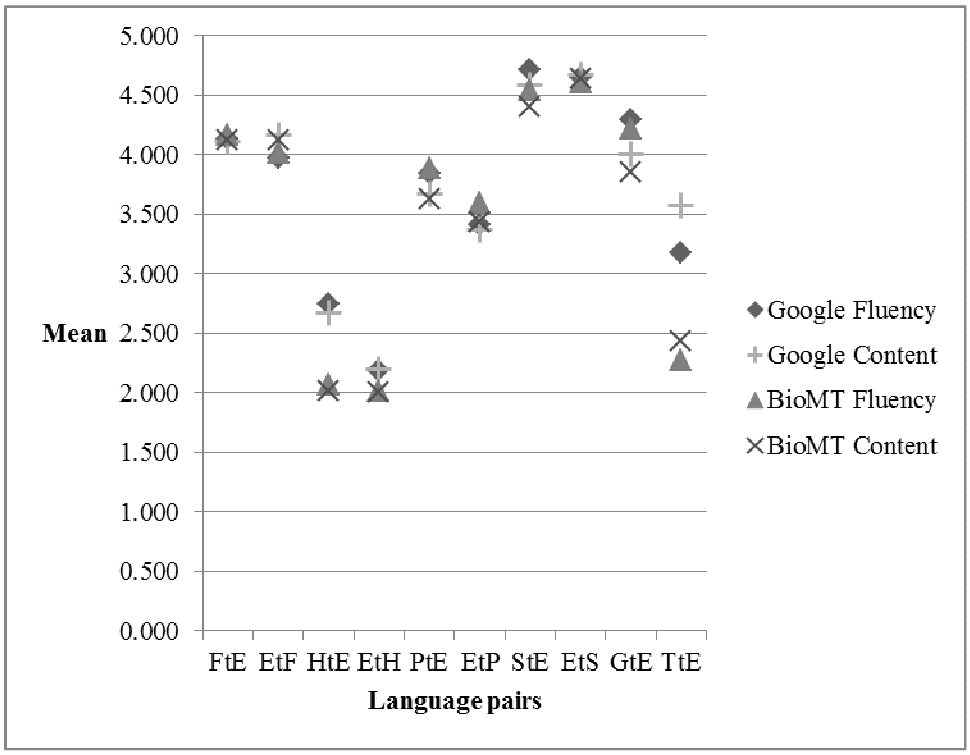

Figure 6. Mean fluency and content scores for the two systems

Table 6 presents the judges' "voting" decisions (SYS1 vs SYS2) in response to the "Which system is better?" question. Bold fonts indicate statistically significant difference by the Chi-square test ($p<0.05$) in favor of a system.

Table 6. Results of the Chi-Square test

	Google (observed N)	BioMT (observed N)	Chi-Square
FtE	104	130	0.089
EtF	33	37	0.633
HtE	**262**	75	0.000
EtH	**115**	71	0.001
PtE	163	148	0.395
EtP	72	83	0.377
StE	**140**	83	0.000
EtS	56	66	0.365
GtE	86	73	0.303
TtE	**162**	29	0.000

Table 7 presents the BLEU scores for each language pair when the BioMT system is trained on all available data but tested only on the same 100 titles that were used for human evaluation.

Table 7. Automated BLEU score generated for the 100 human judged translations

	FtE	EtF	HtE	EtH	PtE	EtP	StE	EtS	GtE	EtG	TtE	EtT
BioMT	44.01	46.73	16.39	8.23	36.24	36.86	53.24	48.89	36.65	34.99	13.68	12.22
Google	36.05	36.05	17.61	7.9	32.54	17.93	48.82	43.91	34.56	21.37	26.41	12.99

In summary, the BioMT system achieved numerically higher BLEU scores in case of nine language pairs while GoogleTranslate had numerically higher scores in three cases. Only the Hungarian-English language pair showed split results between opposite directions of translations. The mean value of human judges' decisions was numerically higher for BioMT for "fluency" in four and for "content" in two translation directions. Meanwhile, GoogleTranslate achieved numerically higher "fluency" in six and "content" scores in eight cases. Cumulatively the human judges voted BioMT's translation a better output in four translation directions and GoogleTranslate's in six. Statistical significance tests did not always supported the numerically higher performance findings. Finally, the results indicate that the increasing size of the training corpus continues to improve the performance of the BioMT system as measured by the automated BLEU score.

Discussion

Figures 3 (BLEU scores per languages and systems) and 6 (human judged scores of fluency and content) show good albeit not perfect correlation between BLEU and human judgment across the studied language pairs. (Figure 3 presents two additional scores compared to Figure 6, because we could generate BLEU statistics for English to German and English to Turkish translations while we had no access to bilingual human judges for those translation directions.) These findings corroborate published results from the general-purpose machine translation field, that the BLEU score is a viable automatic measure of translation quality in the biomedical domain, as well.

On the other hand, while higher BLEU scores indicated that BioMT provided better quality translations for most of the translation directions (Figure 3 and Table 2), statistical significance tests of the human judgments did not support this finding. Table 5 illustrates that the judges scored fluency and content more frequently higher for the output of Google Translate than for BioMT. Google Translate was scored higher than BioMT for six translations for fluency and eight translations for content measures. BioMT was scored higher than Google Translate for four translations for fluency and twice for content. The fluency and content scores correlated remarkably well. Only in cases of English to French and Polish to English translations did the judges score across systems (higher fluency scores for BioMT while higher content scores for Google). Only two translation pairs (Hungarian to English and Turkish to English) were statistically significant for the differences between the scores for translation quality for the two systems.

Table 6 shows a mixed picture for the human judgment scores. BioMT was "voted" by the judges four times as the better system (French to English, English to French, English to Polish and English to Spanish). Google was "voted" six times as the system with better translations (Hungarian to English, English to Hungarian, Polish to English, Spanish to English, German to English, and Turkish to English). The Google "wins" were more pronounced (by numeric absolute value) and were found statistically significant (via the Chi-square test at $p<0.05$) in four cases (Hungarian to English, English to Hungarian, Spanish to English, and Turkish to English).

After aligning the three evaluation methods (Figure 3 and Tables 5 and 6) we found that out of the 12 BLEU scores only three did not align (at least partially) with the results of at least one of the human judgment methods (fluency/content or voting for better system output). For English to Hungarian, Spanish to English and German to English translations, the BLEU statistics pointed to the opposite direction than the human judgments. For two translation pairs (English to German and English to Turkish) we do not have human judgment data.

Table 3 and Figures 4 and 5 show that as the size of the training corpora increases so does the BLEU score. This is not surprising as statistical systems tend to do better with larger amount of training data. The data also points to a plateau effect for translation pairs where we had sufficiently large training corpora to experiment meaningfully with the size of the training data (English to French, French to English, German to English, and English to German). However, none of the studied translation pairs "arrived" to the plateau, yet. It is likely that as the parallel corpora accumulate the quality of the machine translation will improve even without further breakthrough in the translation algorithms. This is good news for investigators working on or planning to work on biomedical machine translation systems.

Finally, Table 7 shows the BLEU scores on the small test sets with 100 titles, which is the set used for human judgment. The results correlate exceptionally well with results from large (occasionally 600 times larger) test sets, and they allow us to compare BLEU and human judgment on the same test sets.

It is noteworthy that building an in-house high performance statistical machine translation system that produces results comparable to the state of the art Google Translate (for translating PubMed titles), according to both human judgment and automated BLEU measurements, is relatively straightforward. All the "parts" necessary to build the system are available as open source components. The parallel corpora are also leasable (free of charge) from the National Library of Medicine. Compared to using an off-the-shelf translation such as Google Translate, which is a black box to the public, the advantages of having an in-house machine translation system are enormous. The in-house system is trained by in-domain data (PubMed) titles, it can be re-trained when more training data become available (which is the case as the number of PubMed titles increases over time), and more training data will result in better translation performance, as shown in Table 3. In addition, maintenance of the in-house system requires minimal or no effort as both the collection of the accumulating parallel corpora and the retraining of the system is easy to automate.

Some of the limitations of our research include that we did not have the same parallel corpora across languages. This makes it impossible to compare translation quality across the studied languages. A second limitation is that the judges were untrained for scoring translation output. However, this limitation is somewhat mitigated by the fact that the translation outputs are intended for "untrained" users (e.g. patients who do not speak English) and if a future version of BioMT will be deployed then its output will be read and interpreted by "untrained" users. In future research we will address the limitations mentioned. We will also develop post-processing steps specific for the biomedical domain to enhance the quality of the translations. We plan to explore the capabilities of the in-house built translation engine to translate PubMed abstracts, in addition to titles.

Conclusion

In answering our research question, we conclude that "We are almost there for some languages but very far for others". For languages (German, Spanish and French) with large training corpora already accumulated in PubMed, translating the titles with high quality machine translation is almost a reality. For these languages the average fluency and content (human judgment) scores were all above four on a five-point scale and in case of Spanish-English and English-Spanish translations the mean scores were very close to the maximum. For languages with small training corpora, the translation quality was very low. Based on the BLEU statistics we conclude that at the present state of statistical machine translation -- in order to generate high quality translations -- the study language needs a training corpus with at least 100K lines of parallel reference titles. Furthermore, the results support BLEU as a viable machine translation evaluation approach in the biomedical domain.

Acknowledgement

Kristina Toutanova from Microsoft Research provided invaluable comments for a pilot version of this work. Dina Demner-Fushman from NLM brought the parallel corpora to our attention. We greatly appreciate their assistance. The project described was supported by Grant Numbers 1K99LM010227-01 and 7R00LM010227-03 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine. The authors would like to thank three anonymous reviewers for their suggestions.

References

1. People Origins and Language. (n.d.). Census Bureau Home Page. Retrieved March 17, 2011, from http://factfinder.census.gov/jsp/saff/SAFFInfo.jsp?_pageId=tp7_origins_language.
2. Hispanic population of the United States. (n.d.). census bureau home page. Retrieved June 10, 2010, from http://www.census.gov/population/www/socdemo/hispanic/hispanic_pop_presentation.html
3. Philipp Koehn, 2010. Statistical Machine Translation, Cambridge University Press, New York.
4. Google Translate. (n.d.). Retrieved March 17, 2011, from http://translate.google.com
5. E.H. Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In Proceedings of the
   Eagles Workshop on Standards and Evaluation, Pisa, Italy.
6. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311-318, Philadelphia, PA, July.
7. Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006), pages 223-231, Cambridge, MA, August.
8. Alon Lavie and Michael Denkowski, "The METEOR Metric for Automatic Evaluation of Machine Translation", Machine Translation, 2010.
9. Fang Liu, Paul Fontelo and Michael Ackerman. 2006. BabelMeSH: Development of a cross-language tool for Medline/PubMed. AMIA Annual Symposium Proceedings, 2006: 1012.
10. Improving access to multi-lingual health information through machine translation. (n.d.). Project Information NIH Research Portfolio Online Reporting Tools. Retrieved March 17, 2011, from http://projectreporter.nih.gov/project_info_description.cfm?aid=7946175&icde=7444342
11. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic.
12. Peter Brown, Vincent Pietra, Stephen Pietra and Robert Mercer, 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2), pages 263-311.
13. Andreas Stolcke. 2002. SRLIM – An Extensible Language Modeling Toolkit. In Proc. of the International Conference on Spoken Language Processing, vol 2, pages 901-904. Denver, USA.
14. Google-api-translate-java - project hosting on Google Code. (n.d.). Google Code. Retrieved June 10, 2010, from http://code.google.com/p/google-api-translate-java/
15. Solomon FM, Eberl-Lefko AC, Michaels M, Macario E, Tesauro G, Rowland JH. Development of a linguistically and culturally appropriate booklet for Latino cancer survivors: lessons learned. Health Promotion Practice, 2005 Oct;6(4):405-13.