Enabling Monolingual Translators: Post-Editing vs. Options

Philipp Koehn
University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB
Scotland, United Kingdom
pkoehn@inf.ed.ac.uk

Abstract

We carried out a study on monolingual translators with no knowledge of the source language, but aided by post-editing and the display of translation options. On Arabic-English and Chinese-English, using standard test data and current statistical machine translation systems, 10 monolingual translators were able to translate 35% of Arabic and 28% of Chinese sentences correctly on average, with some of the participants coming close to professional bilingual performance on some of the documents.

While machine translation systems have advanced greatly over the last decade, nobody seriously expects human-level performance any time soon, except for very constraint settings. But are todays systems good enough to enable monolingual speakers of the target language without knowledge of the source language to generate correct translations? And what type of assistance from machine translation is most helpful for such translators?

We carried out a study that involved monolingual translators who had no knowledge of Chinese and Arabic to translate documents from the NIST 2008[1] test sets, being assisted by statistical machine translation systems trained on data created under the GALE[2] research program.

Our study shows that monolingual translators were able to translate 35% of Arabic and 28% of Chinese sentences, under a strict standard of correctness that scored professional bilingual translations as 61% and 66% correct for Arabic and Chinese, respectively. We found also large variability among the participants and between the documents in the

---

[1]http://www.itl.nist.gov/iad/mig/tests/mt/
[2]http://www.darpa.mil/ipto/programs/gale/gale.asp

study, indicating the importance of general language skills and domain knowledge. The results suggest that a skilled monolingual translator can compete with a bilingual translator, when using todays machine translation systems.

1 Related Work

The use of human translators in combination with machine translation is as old as the emergence of the first effective machine translation systems. Typically, this takes the form of a human translator post-editing machine translation output, and rarely of a human translator guiding the decisions of a machine translation system. Recent examples of using post-editing of machine translation in tools for translation tools are the Google Translator Toolkit (Galvez and Bhansali, 2009) and the WikiBabel project (Kumaran et al., 2008).

A recent seminal effort on building interactive machine translation systems (Langlais et al., 2000; Barrachina et al., 2009) looked at a tighter integration of machine translation and human translation by developing a prediction model that interactively suggests translations to the human translator, taking her prior translation decisions into account. This approach was recently re-implemented and extended by Koehn (2009).

Our study uses both post-editing and the extended interactive machine translation approach as types of assistance for translations. In our case, however, we look at monolingual translators, while prior work has focused on bilingual translators.

Another effort to enable monolingual translators looked at a more linguistically motivated tool using syntactic analysis to inform their translation decisions (Albrecht et al., 2009).

The quality of the translations produced by

monolingual translators was previously explored by Callison-Burch (2005) in a submission to the NIST 2005 evaluation campaign, but not properly evaluated. The idea of using post-editing by monolingual speakers without access to the source as a metric to evaluate machine translation quality of different systems was explored by Callison-Burch et al. (2009) in the WMT 2009 shared task.

2 Human Translation

Except for constraint settings with a very limited domain, translation quality by trained humans is much higher than automatic translation methods. Especially for the commercially most relevant field of publication-quality translation of official reports, product manuals, promotion material, web sites, and so on, machine translation currently plays at most a supportive role.

2.1 Translation Tools

The main draw-back of relying on professional human translators is their high cost. A number of technological advances in the industry have increased the productivity of translators, and thus lowered their cost, over the last two decades. The pervasive use of computers and the Internet has reduced the cost of management, and helped a industry where translation is outsources many times over: from the original customer to a translation agency, from a translation agency to freelance translators, and maybe some additional levels in between.

The use of computers has also led to the adoption of tools such as translation memories[3] (databases of translated material that are queried for fuzzy matches, i.e. translated sentences similar to the one to be processed), monolingual and bilingual concordances (showing words used in context, and their translations), terminology databases, online dictionaries and thesauri, and basic editing tools such as word processors and spell checkers (Desilets, 2009).

The use of machine translation has not yet made great inroads into the toolbox of professional translators. Being reduced to mere post-editors of badly machine translated texts is not an appealing prospect, and machine translation is generally considered (rightly or wrongly) not yet good enough to

increase productivity. More innovative use of machine translations such as interactive machine translation (Langlais et al., 2000) has not advanced much beyond the research stage. There is rich potential for improvements and entirely new tools.

2.2 Translation Skills

A fully qualified professional translator has to have two sets of skills when translating a text. On the one hand the language skills to generally understand the source language and to write well in the target language, and on the other hand the domain knowledge to understand the content of a possibly very specialized technical document. Both skill sets may be hard to find, especially in combination.

In fact, it is common practice in the translation industry to differentiate translators according to their qualifications. For instance, junior translators may produce the first draft, and senior translators edit it — which they will be able to do much faster than a translation from scratch by themselves.

Human translation is also performed in a non-professional environment by generally less qualified volunteer translators. To give just a few example: there are vibrant communities that concern themselves with the translation of Wikipedia articles[4] (Kumaran et al., 2008), open source software documentation,[5] movie subtitles,[6] and even material such as the TED conference talks.[7]

Research has shown that less qualified translators are able to increase their productivity and quality disproportionally when given automatic assistance (Koehn and Haddow, 2009). Assistance may be as limited as offering machine translation in a post-editing environment, as for instance provided by Google Translator Toolkit[8] (Galvez and Bhansali, 2009) which provides a special function to translate Wikipedia articles.

In this context, our work looks at one extreme of the skill range: translators that have no knowledge of the source language. While we would not expect them to compete with professional translators that have this knowledge, their inferior performance may

---

[3] for instance: Trados, http://www.trados.com/

[4] http://en.wikipedia.org/wiki/Wikipedia:Translation
[5] http://l10n.kde.org/
[6] http://www.opensubtitles.org/
[7] http://www.ted.com/translate/
[8] http://translate.google.com/toolkit/

Figure 1: Translation Options shown to the monolingual translator. The machine translation of the Arabic input sentence is: *The us house of representatives adopted thursday a law calls for the withdrawal of us combat troops from iraq by the first of april 2008, defying once again president george bush who opposed to setting any date.*

be remedied as suggested above: their texts may be edited by a more qualified translator, or their domain knowledge may augment the language skills of a collaborating translator.

From the view of human translation, the main question that this paper is trying to answer is: how well can monolingual translators perform, given the current quality of machine translation, and what types of assistance offered by a machine translation system is most helpful?

3 Machine Translation

Statistical machine translation has made great progress over the last two decades, with changing models, learning methods, decoding algorithms, decision rules, etc. While there is increasing effort to build grammar-based translation models that take into account the recursive nature of language, currently the most popular models are still phrase-based.

3.1 Phrase-Based Models

In phrase based models, the input is segmented into text chunks (that do not have to correspond to linguistic phrases), each is translated and may be reordered, and the output is assembled with the help of a language model. The translations for individual phrases are called translation options. Typically, up to 20 translation options for each input phrase are considered during decoding.

The large number of translation options and their even larger combinatorial arrangement creates a search space that is too large to exhaustively explore, creating the need for a heuristic search algorithm. During the heuristic search a search graph is constructed. This search graph can be converted into a word lattice, which is useful for n-best list generation, or in our case interactive machine translation.

3.2 Interactive Machine Translation

The by-products of phrase-based models have been used in a type of computer aided translation tool called interactive machine translation. In these setups, the human translator is creating the translation, but receives suggestions how to complete the sentence or is offered alternative translation options for the input words and phrases.

The translation options that the decoder is using are ranked based on their probability and presented to the human translator, as done by Koehn (2009). Sentence completion prediction is based on the search graph (Barrachina et al., 2009). If the human translator starts a translation that diverges from the suggestion, the interactive translation tool quickly computes a approximate match in the search graph and uses this as a starting point for further predictions.

In our experiments, we offer both translation options and interactive sentence completion predictions to the user. See Figure 1 for an example.

539

3.3 Arabic and Chinese

This paper is using machine translation systems for Arabic–English and Chinese–English that were developed in the context of the recent GALE research program funded by DARPA.

The choice of these two languages pairs has two motivations: first, a lot of resources have gone into improving translation quality for these language pairs. An important question is how the improvements in translation quality can be utilized.

The second motivation for choosing Arabic–English and Chinese–English is that they are undecipherably foreign for a typical European or American speaker of English. The fact that both languages are written in a different script already makes it impossible to spot cognates, except for the occasional number. In our study, the test subjects had to practically exclusively rely on the given sentence translations or phrase translations options.

3.4 Evaluation of Machine Translation

Chinese–English is considered significantly harder than Arabic–English, as measured by automatic metrics (which measure similarity to a human reference translation), human evaluation metrics such as HTER (which measures the number of editing steps necessary to correct the output into an acceptable translation), or human judgment on the correctness of the translation, its fluency or adequacy (which is typically measured on a scale from 1 to 5).

All these metrics have been criticized in the past as too simple, biased towards statistical systems, non-repeatable, having low intra and inter-annotator agreement, or plainly too expensive to use. How to properly evaluate machine translation quality is still an open problem.

From the view of machine translation evaluation, this paper explores the question if current machine translation systems have reached the goal to bring across the meaning of a foreign text. The ability of a monolingual target language speaker to produce correct translations (based on her understanding of the machine translation output) is a test for this goal. It sets aside the problems of clumsy wordings and grammatical errors. To relate this to traditional error metrics in machine translation: we focus on the adequacy opposed to the fluency of translation.

Language	Sentences	Words
Arabic	9,320,356	228,712,189
Chinese	2,039,399	49,564,193

Table 1: Training data: number of sentences and English words in the parallel training data

4 Experiment

We trained translation models using Moses (Koehn et al., 2007) on the bilingual data provided by the LDC, with additional monolingual data from the English Gigaword corpus for an interpolated 5-gram language model. Basic statistics about the corpus are given in Table 1. The systems are close to the state of the art.

We used four news stories for each of the two languages for the monolingual translators. The news stories were selected from the evaluation sets of the 2008 machine translation evaluation campaign organized by NIST. See Table 2 for details. The news stories are rather short (around 10 sentences each), since we opted for a variety of stories rather than long stories.

The evaluation set comes with four reference translations. This allowed us to use one of the reference translation as gold standard for the evaluation, and the other three reference translations as competitors for the monolingual translations.

We recruited 10 monolingual translators, students at the University of Edinburgh for the study. None of the students had knowledge of either Chinese or Arabic. Each translator was given all eight stories to translate, half of the stories with only the machine translation output (Post-editing task) and half of the stories with interactive assistance as described in Section 3.2: prediction of sentence completion and translation options (Options).

In both cases, we also displayed the Arabic or Chinese source sentence to the translator, which may show some clues regarding punctuation, numbers, or the length of source words. The translators had no knowledge of the source script.

After all the translations were completed, we assessed the translation quality. Since we did not have access to bilingual speakers for this, we resorted to the standard manual setup, where human judges are asked to assess the quality of each sentence transla-

Story	Headline	Sent.	Words
1: Chinese	White House Pushes for Nuclear Inspectors to Be Sent as Soon as Possible to Monitor North Korea's Closure of Its Nuclear Reactors	6	207
2: Chinese	Torrential Rains Hit Western India, 43 People Dead	10	204
3: Chinese	Research Shows a Link between Arrhythmia and Two Forms of Genetic Variation	7	247
4: Chinese	Veteran US Goalkeeper Keller May Retire after America's Cup	10	367
5: Arabic	Britain: Arrests in Several Cities and Explosion of Suspicious Car	7	224
6: Arabic	Ban Ki-Moon Withdraws His Report on the Sahara after Controversy Surrounding Its Content	8	310
7: Arabic	Pakistani Opposition Leaders Call on Musharraf to Resign.	11	312
8: Arabic	Al-Maliki: Iraqi Forces Are Capable of Taking Over the Security Dossier Any Time They Want	8	255

Table 2: News stories used in the experiment with headlines from the reference translation

Assistance	Arabic	Chinese
Bilingual	61±6%	66±6%
Postediting	35±4%	26±4%
Options	34±4%	30±4%

Table 3: Correctness of translations (with 95% confidence interval) under the two types of assistance, compared against professional reference translations

Translator	Arabic	Chinese
bi1	67±10%	65±11%
bi2	49±10%	67±10%
bi3	67±10%	67±9%
mono1	48±11%	31±11%
mono2	29±10%	21±8%
mono3	26±10%	12±7%
mono4	50±11%	26±11%
mono5	25±10%	25±10%
mono6	26±9%	18±9%
mono7	23±10%	29±10%
mono8	50±11%	50±10%
mono9	42±10%	37±11%
mono10	25±9%	32±10%

Table 4: Correctness by translator (note: different bilingual translators for Arabic and Chinese)

tion compared to a reference translation in context — the first reference translation in the NIST evaluation set which was produced by a professional translation agency.

We used a strict evaluation metric: a binary judgment, if the translation is correct. Correct was defined as *a fluent translation that contains the same meaning in the document context*. The reference translation was shown with its document context (two sentences before and after). We used a variant of the web-based evaluation tool of the 2009 Workshop on Statistical Machine Translation.

5   Results

The headline results are displayed in Table 3. The bilingual translations which were taken from the other three reference sets score surprisingly low: only about two thirds of the sentences were deemed to be correct by our judges. This is a better result than the monolingual translators performance, who translate around one third of the sentences correctly, except for a statistically significant worse showing for post-editing Chinese–English.

Translation speed of the monolingual translators varied, but it was mostly around 500 words per hour (7 seconds per word), which is roughly comparable to the lower end of professional translation speed.

Table 4 shows the performance of the individual translators. The 95% confidence intervals are very wide, due to the few sentences that were translated by each translator, but some monolingual translators are significantly better than others. Some of the monolingual translators seem to compete head-to-head with the professional bilingual translators: three monolingual translators perform as well as one of the bilingual translators for Arabic–English, albeit one has to be cautioned by the wide confidence intervals. See also Figure 2 for a graphical display.

Story	BLEU	Bilingual	Post-ed.	Options
1: Chinese	42.8	76±16%	32±13%	40±13%
2: Chinese	24.8	70±10%	39±8%	33±9%
3: Chinese	35.1	61±12%	19±8%	17±7%
4: Chinese	26.7	64±11%	12±6%	36±9%
5: Arabic	43.6	60±14%	10±6%	13±7%
6: Arabic	48.5	57±13%	34±9%	43±9%
7: Arabic	60.5	72±10%	45±8%	36±9%
8: Arabic	55.7	50±13%	45±10%	39±10%

Table 5: Correctness by story and BLEU score of MT

	Length	Bilingual	Post-ed.	Options
Arabic	≤15 words	81±16%	56±15%	48±16%
Arabic	16–30 words	54±10%	41±8%	37±7%
Arabic	>30 words	62±8%	27±6%	29±6%
Chinese	≤15 words	60±12%	48±10%	21±9%
Chinese	16–30 words	73±13%	25±9%	32±10%
Chinese	>30 words	68±8%	17±5%	33±6%

Table 6: Correctness by sentence length

Similarly, performance on the different stories varies (Table 5, Figure 3): For instance, the monolingual translators struggled with the Chinese medical and sports stories (no. 3 and 4) and the Arabic car explosion story (no. 5), while even on average, they are close to bilingual translation quality on the Arabic stories 6 and 8. Note that correctness correlates mildly with BLEU.

Surprisingly, we did not find a consistent effect of sentence length on the quality of the translations (see Table 6). We expected to find worse translations among the longer sentences, but this is not the case for the all conditions.

6 Analysis

Our results have shown that monolingual translators are often able to produce correct translation when post-editing output from current Arabic–English and Chinese–English machine translation systems. For Chinese–English, they are better when given additional assistance in form of translation options and interactive machine translation.

We give in Figure 4 examples for translations by machine translation, as well as monolingual and bilingual translators.

One puzzle is the low score for the professional human translators, as only two thirds of their trans-

(a) Critical judges

REF: *Torrential Rains Hit Western India, 43 People Dead*

BI: *Heavy Rains Plague Western India Leaving 43 Dead*

(b) Mistakes by the professional translators

REF: *Over just two days on the 29th and 30th, rainfall in Mumbai reached 243 mm.*

BI: *The rainfall in Mumbai had reached 243 cm over the two days of the 29th and 30th alone.*

(c) Bad English by monolingual translators

MONO: *The western region of india heavy rain killed 43 people.*

(d) Mistranslated / untranslated name

REF: *Johndroe said that the two leaders ...*

MT: *Strong zhuo, pointing out that the two presidents ...*

MONO: *Qiang Zhuo pointed out that the two presidents ...*

(e) Wrong relationship between entities

REF: *The next match against Colombia will probably be the US team's and Keller's last performance in this America's Cup competition.*

MT: *The colombian team for the match, and it is very likely that the united states and kai in the americas cup final performance.*

MONO6: *The Colombian team and the United States are very likely to end up in the Americas Cup as the final performance.*

MONO8: *The next match against Colombia is likely to be the United States' and Keller's final performance in the current Copa America.*

(f) Badly muddled machine translation

REF: *In the current America's cup, he has, just as before, been given an important job to do by head coach Bradley, but he clearly cannot win the match singlehanded. The US team, made up of "young guards,"...*

MT: *He is still being head coach bradley appointed to important, it's even a fist ", four young guards at the beginning of the ", the united states is...*

MONO: *He is still being considered important by head coach Bradley who appointed him. It is a fight with "four young guards at the beginning of their careers", but the United States...*

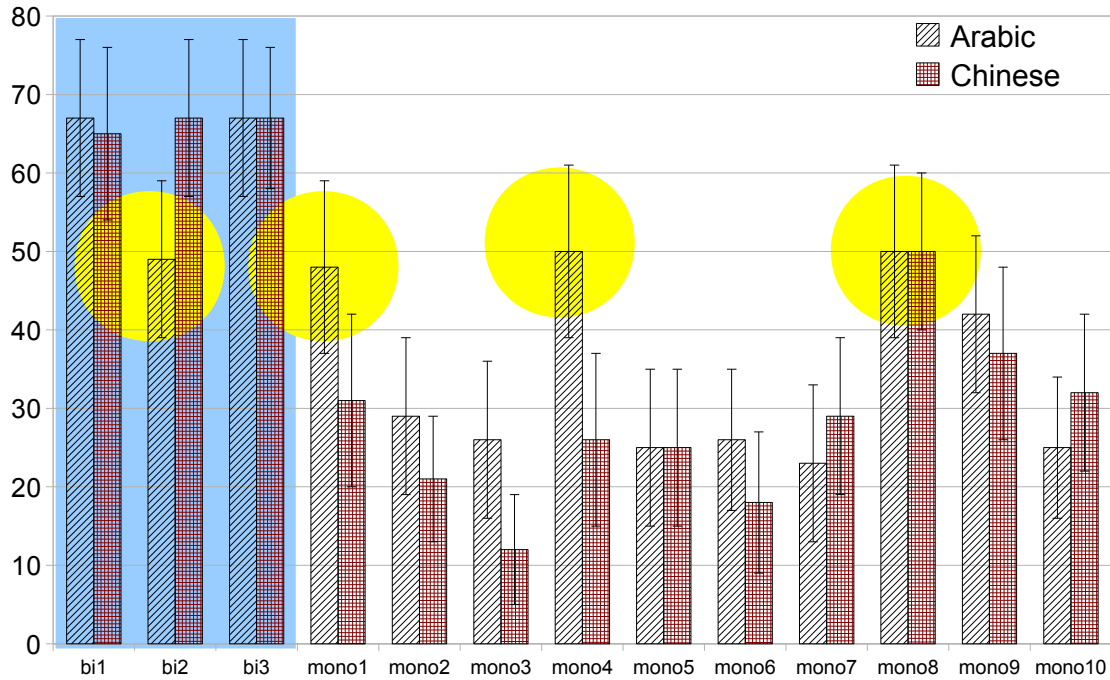Figure 4: Examples of translations

Figure 2: Quality of different bilingual and monolingual translators: For Arabic, three monolingual translators are as good as the worst bilingual translator (around 50% of sentences judged as correct). For detailed numbers, see Table 4.
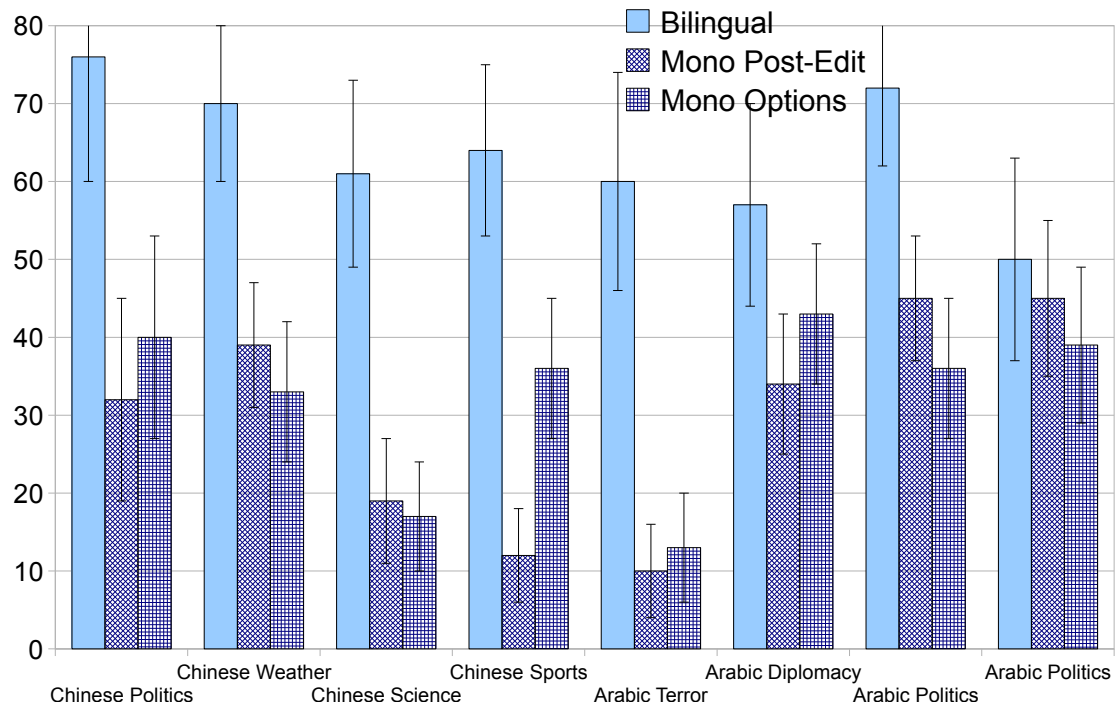

Figure 3: Translation quality of monolingual translators differs significantly between stories: For the last Ararbic politics stories average performance is close to bilingual quality, while it is bad for the Chinese science and sports as well as the Arabic terror story. For detailed numbers, please see Table 5.

lations were deemed to be correct. The example (a) shows such a translation, and it is hard to tell why it was deemed wrong by all three judges who looked at it. There are real mistakes in the professional translations, as example (b) shows, which mistakes the rain fall amount as *243cm* instead of *243mm*.

Some monolingual translators, by the way, also had problems with that number. The machine translation system is not very well in translating numbers, which could be relatively easily addressed.

Sometimes monolingual translators are just not thorough enough in their efforts, as example (c) shows, where the output does have the correct meaning elements, but it is just not correct English. These type of examples explain the big difference between the different monolingual translators.

A severe problem for monolingual translators are untranslated or mistranslated names. In example (d) *Johndroe* was referred to by monolingual translators as *Qiang Zhuo* or *Strong Zhuo*. The statistical machine translation system we used has no special name transliteration component, so often a name remains untranslated. Without given the right translation as a choice, the monolingual is in no position of completing a correct translation.

The monolingual translators' world and domain knowledge helps them a great deal to piece together translations, but sometimes it is not enough, as example (e) shows. There is some connection between *final performance*, *United States* and *Columbia*, but it is not the final performance for both teams as MONO6 renders it. Translator MONO8 got it right, but other translators made different mistakes.

Finally, there are some cases, as example (f) shows, where the machine translation is just so bad, that monolingual translators have no chance to render a proper translation of the sentence, especially when only post-editing.

7 Conclusion

We approached this study from two directions: the motivation to enable monolingual translators and the need for a way to assess the quality of todays machine translation systems.

Coming from a human translator's perspective, we asked what type of assistance machine translation can provide for a human translator. We compared the use of interactive machine translation against post-editing, and found no significant difference for Arabic (34% vs. 35%), but better performance with richer assistance for Chinese (30% vs. 26%). We believe that there is ample opportunity to provide additional assistance and we will explore this in future work.

Coming from a machine translation perspective, we asked if current systems are good enough to bring across the meaning of documents, even if generating output language with grammatical and idiomatic mistakes. Given the harsh metric we use to assess translation quality (complete correctness of a sentence), we showed that monolingual translators were able to produce translations that were on average 35% (Arabic) and 28% (Chinese) correct, compared to 61% (Arabic) and 66% (Chinese) correctness for professional bilingual translations.

Arguable, the method we use to assess the preservation of meaning in machine translation is superior to subjective adequacy judgments: it separates the task of defining the meaning of a machine translation from the assessment of its correctness.

We identified name and number translation as important aspects to improve performance on this task.

We also learned that there are significant difference between human translators, which indicates that general language skills and effort are very important. We also learned that the performance varies significantly for different documents in a way that hints at the importance of domain knowledge. In conclusion, a good monolingual translator has good language skills in the target language and understands the domain. In this case, this study suggests, she may be as good as a professional bilingual translator.

Acknowledgement

References

Albrecht, J., Hwa, R., and Marai, G. E. (2009). Correcting automatic translations through collaborations between MT and monolingual target-language users. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 60–68, Athens, Greece. Association for Computational Linguistics.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Callison-Burch, C. (2005). Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of Machine Translation Evaluation Workshop*.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Desilets, A. (2009). Up close and personal with a translator - how translators really work. In *Machine Translation Summit XII*. Tutorial.

Galvez, M. and Bhansali, S. (2009). Translating the world's information with google translator toolkit.

Koehn, P. (2009). A web-based interactive computer aided translation tool. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*.

Koehn, P. and Haddow, B. (2009). Interactive assistance to human translators using statistical machine translation methods. In *Machine Translation Summit XII*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (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. Association for Computational Linguistics.

Kumaran, A., Saravanan, K., and Maurice, S. (2008). wikiBABEL; community creation of multilingual data. In *Babel Wiki Workshop 2008: Cross-Language Communication*.

Langlais, P., Foster, G., and Lapalme, G. (2000). Transtype: a computer-aided translation typing system. In *Proceedings of the ANLP-NAACL 2000 Workshop on Embedded Machine Translation Systems*.