Using Monolingual Human Computation to Improve Language Translation via Targeted Paraphrase

Philip Resnik
Linguistics and UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

Chang Hu
Computer Science and HCIL
University of Maryland
College Park, MD
changhu@cs.umd.edu

Olivia Buzek
Linguistics and Computer Science
University of Maryland
College Park, MD
olivia.buzek@gmail.com

Benjamin B. Bederson
Computer Science and HCIL
Human Computer Interaction Laboratory
University of Maryland
College Park, MD
resnik@umd.edu

ABSTRACT

We introduce a new approach to the problem of obtaining cost-effective, reasonable quality translation, by exploiting simple and inexpensive human computations by monolingual speakers. The key insight behind the process is that it is possible to to spot likely translation errors with only monolingual knowledge of the target language, and it is possible to generate new ways to say the same thing (i.e. paraphrases) with only monolingual knowledge of the source language. Initial evaluation demonstrates substantial improvements in translation quality.

1. INTRODUCTION

Consider the following common scenario. You have some text in another language, which you can't read, and you'd like to translate it into your own language. You run it through {Google Translate, Babelfish, Babylon}, and the result is... well, you still have text you can't read. There are pieces you can make sense of, and pieces you can't. What do you do now?

At this point, the options are fairly limited. A human needs to be brought into the loop in one way or another. If it's important enough, perhaps you place an order with an online translation service, at a typical cost of some \$0.20-\$0.25 per word, and wait a typical minimum of a day or so for the result. If it's moderately important and you've got the technical skills, perhaps you create one or more HITs on Amazon's Mechanical Turk to obtain a reasonable translation a little more quickly and at lower cost — *if* you're lucky enough to locate competent translators for the language pair. This is fairly likely if translating from more widely spoken languages into English. It's a a good deal less likely for virtually any other language pair, e.g. translating from Farsi into French. (That's a problem for professional translation, too.) If it's less than moderately important, chances are you give up.

In this paper, we propose a new approach to the translation problem that focuses on a relatively unexplored part of the cost/quality spectrum: utilizing speakers of the source and target languages who are *effectively monolingual*, i.e. who each only know one of the two languages relevant for this translation task, working in tandem with machine translation (MT) to identify and fix problems in automatic translations. The solution we are proposing provides a more cost-effective approach to translation in scenarios where machine translation *would* be considered acceptable to use, if it were generally of high enough quality. This would clearly exclude tasks like translation of medical reports or business contracts, where there the validation of a qualified bilingual translator is absolutely necessary. However, it does include many real-world scenarios, such as following news reports in another country, gauging international opinion about a product, or generating a decent first draft translation of a Wikipedia page for Wikipedia editors to improve.

Post-editing after previous translation, especially MT, has long been a part of translation, e.g. [10]. Callison-Burch et al. [4] pioneered the exploration of *monolingual* post-editing within the MT community, an approach extended more recently to provide richer information to the user by Albrecht et al. [1] and Koehn [9]. There have also been at least two independently developed human-machine translation frameworks that employ an iterative protocol involving monolinguals on both the source and target side. Morita and Ishida [11] describe a system in which target and source language speakers perform editing of MT output to improve fluency and adequacy, respectively; they utilize source-side paraphrasing at a course grain level, although their approach is limited to requests to paraphrase the entire sentence when the translation cannot be understood. Hu et al. [2] describe

HCOMP 2010, Washington DC

a similar protocol in which cross-language communication is enhanced by metalinguistic communication in the user interface. Shahaf and Horvitz [12] use machine translation as a specific instance of a general game-based framework for combining a range of machine and human capabilities.

The technique we describe here can be viewed as compatible with the richer protocol- and game-based approaches, but it is considerably simpler. We demonstrate how monolingual human computation can be used to improve translation by means of *human-targeted paraphrasing*, in which target-language monolinguals identify parts of the translation that don't appear to be right, and source-language monolinguals provide the MT system with alternative phrasings that might lead to better translations. This is an alternative variant to our recent exploration of *machine*-targeted paraphrasing [3], in which the targeting is done automatically; in that exploration we demonstrated, via an oracle study, that quite substantial improvements in translation quality are available when source-language monolinguals provide alternative phrasing for parts of the sentence that are heuristically chosen as likely to be poorly translated.

In Section 2, we describe how our targeted paraphrase translation process works. In Section 3 we apply this approach to a small study of translation of sentences from Chinese Wikipedia into English. In Section 4 we discuss the results, which demonstrate substantial improvements in translation quality. In Section 5 we present conclusions and discuss future work.

2. TARGETED PARAPHRASE

The basis for our approach is the observation that the source sentence provided as input to an MT system is just one of many ways in which the meaning could have been expressed, and that for any given MT system some forms of expression are easier to translate than others. This observation has been applied quite fruitfully over the past several years to deal with challenges involving segmentation, morphological analysis, and more recently source langage word order [5, 6, 7]. Here we apply it to the surface expression of meaning.

For example, consider the following translation from English to French by an automatic MT system:

- **Source:** Polls indicate Brown, a state senator, and Coakley, Massachusetts' Attorney General, are locked in a virtual tie to fill the late Sen. Ted Kennedy's Senate seat.
- **System:** Les sondages indiquent Brown, un sénateur d'état, et Coakley, Massachusetts' Procureur général, sont enfermés dans une cravate virtuel à remplir le regretté sénateur Ted Kennedy's siège au Sénat.

A French speaker (indeed, even someone with even a semester of college French) can look at this automatic translation and see that the underlined parts are probably wrong. If we change the source sentence to rephrase the underlined pieces (e.g. changing *Massachusetts' Attorney General* to *the Attorney General of Massachusetts*), and then use the same MT system again, we obtain a translation that is still imperfect, but is more acceptable:

- **System:** Les sondages indiquent que Brown, un sénateur d'état, et Coakley, le procureur général du Massachusetts, sont enfermés dans une cravate virtuel ÃăpourvoirlesiÃÍgeauSénatdeSen.TedKennedy,quiestdécédérécemment.

Our system makes this general idea operational as follows.

Initial machine translation..

In our system we use Google Translate. For this paper, our source sentences are in Chinese, with English as the target.

Human computation: error identification..

Turkers are instructed to identify parts of the English sentence that are ungrammatical, nonsensical, or apparently incorrect.

Projection of error spans back to source..

Using word alignments between the source and translation, the marked spans are projected back to identify the corresponding Chinese source spans that generated them.

Human computation: targeted paraphrase..

A Turker sees a Chinese sentence with a phrase marked, and is asked to replace the marked text with a different way of saying the same thing, so that the resulting sentence still makes sense and means the same thing as the original sentence. To illustrate in English, the Turker might see *John and Mary took a European vacation this summer* and supply the paraphrase *Mary went on a European*, verifying that the resulting *John and Mary went on a European vacation this summer* preserves the original meaning.

Generating sentential source paraphrases..

For each sentence, there may be multiple paraphrased spans. These are multiplied out to provide full-sentence paraphrases. For example, if two spans are each paraphrased three ways, there would be six Chinese paraphrases.

Machine translation of alternative sentences..

The alternative Chinese sentences are sent through the same MT system.

At this point, our notional interface would present the new translation alternatives to the English recipient. In our evaluation, we consider the quality of all the alternatives, and we also simulate a process in which the English speaker selects a translation based solely on English fluency, assessing the quality of the selected result.

3. TRANSLATION EXPERIMENT

As data for this experiment, eleven sentences in simplified Chinese were selected from the article on "Water" in Chinese Wikipedia (http://zh.wikipedia.org/zh-cn/%E6%B0%B4). This article was chosen because its topic is well-known in both English-speaking and Chinese-speaking populations. The first five sentences were taken from the first paragraph of the article. The other six sentences were taken from a randomly-chosen paragraph in the article. As a preprocessing step, we removed any parenthetical items from the input sentences, e.g. "(H20)". The shortest sentence in this set has 12 Chinese characters, the longest has 54.[1]

[1]Note that this page is *not* a translation of the corresponding English Wikipedia page or vice versa.

Both error identification HITs and paraphrasing HITs were priced at $0.01. The error identification HITs expire in an hour, whereas paraphrase HITs expire in a day. Each error identification HIT is handled redundantly by 3 Mechanical Turk workers (Turkers), and each paraphrasing HIT is handled redundantly by 5 Turkers.

Four Turkers participated in the error identification stage (in English), and their HITs took an average of 34 seconds to complete; one was excluded as an obvious cheater. For the paraphrase stage (in Chinese), three Turkers participated, taking under 20 seconds per hit.[2] Total payment to Turkers in this experiment amounted to $0.66: $0.16 for error identification and $0.50 for generation of paraphrases.

4. EVALUATION OF RESULTS

The original outputs of Google Translate (GT), and the results of our targeted paraphrase translation process (TP), were evaluated according to widely used critera of fluency and adequacy. Fluency ratings were obtained on a 5-point scale from three native English speakers without knowledge of Chinese. Translation adequacy ratings were obtained from three native Chinese spekaers who are also fluent in English; they assessed adequacy of English sentences by comparing the communicated meaning to the Chinese source sentences.

Fluency was rated on the following scale:

1. Unintelligible: nothing or almost nothing of the sentence is comprehensible.
2. Barely intelligible: only a part of the sentence (less than 50%) is understandable.
3. Fairly intelligible: the major part of the sentence passes.
4. Intelligible: all the content of the sentence is comprehensible, but there are errors of style and/or of spelling, or certain words are missing.
5. Very intelligible: all the content of the sentence is comprehensible. There are no mistakes.

Adequacy was rated on the following scale:

1. None of the meaning expressed in the reference sentence is expressed in the sentence.
2. Little of the reference sentence meaning is expressed in the sentence.
3. Much of the reference sentence meaning is expressed in the sentence.
4. Most of the reference sentence meaning is expressed in the sentence.
5. All meaning expressed in the reference sentence appears in the sentence.

[2]The four English-speaking Turkers were recruited through the normal Mechanical Turk mechanism. The three Chinese-speaking Turkers were recruited offline by the authors in order to quickly obtain results, although they participated as full-fledged Turkers; we will replicate and extend the study for the final version using only Turkers who come to our HITs through the normal Mechanical Turk channels.

For each GT output, we averaged across the ratings of the alternative TP to produce average TP fluency and adequacy scores. The average GT output ratings, measuring the pure machine translation baseline, were 2.36 for fluency and 2.91 for adequacy. Averaging across the TP outputs, these rose to 3.32 and 3.49, respectively.

A more sensible evaluation might not be to *average* across alternative TP outputs, but to simulate the behavior of a target-language speaker who simply chooses the one translation among the alternatives that seems most fluent. If we select the most fluent TP output for each source sentence, according to the English-speakers' average fluency ratings, we obtain average test set ratings of 3.58 for fluency and 3.73 for adequacy. Those are respective gains of 0.82 and 1.21 over baseline MT output, each on a 5-point scale.

Figure 1 shows a selection of outputs: we present the two cases where the most fluent TP alternative shows the greatest gain in average fluency rating (best gain +2.67); two cases near the median gain in average fluency (median +1); and the worst two cases with respect to effect on average fluency rating (worst -0.33).[3] The table accurately conveys a qualitative impression corresponding to the quantitative results: the overall quality of translations appears to be improved by our process rather consistently, despite the absence of any bilingual input in the improvements.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced a new approach to the problem of obtaining cost-effective, reasonable quality translation, by exploiting simple and inexpensive human computations by monolingual speakers. The key insight behind the process is that it is possible to to spot likely translation errors with only monolingual knowledge of the target language, and it is possible to generate new ways to say the same thing (i.e. paraphrases) with only monolingual knowledge of the source language. Although our study thus far has been conducted only on a small scale, it seems evident that judicious use of these human capabilities, in tandem with machine translation, can lead to dramatically improved translation output even when no bilingual speaker is involved in the process.

We are in the process of constructing a user interface to support this approach as a real-world experimental testbed; we are designing the UI, for ease of use, to remain substantially similar to the user experience in Google Translate. We are also continuing to explore the research path in which likely mistranslations are identified automatically [3], as well as the use of an MT system that can take advantage of combining all the human-supplied paraphrases into a single compact representation [5, 6, 7]. We also plan to explore the application of our approach in scenarios involving less-common languages by using a more common language as a pivot or bridge [8].

Finally, we are exploring the applicability of this new concept to the production of fully reliable translations, by using the output of our new process to create significantly improved input for bilingual post-editing. We hypothesize that our approach will enable a new division of labor in which free computation (MT) and inexpensive human computation (MTurk) take on a more substantial portion of the

[3]We can include source Chinese sentences or correct English references in the final version, if so desired.

Condition	Fluency	Adequacy	Sentence
GT	1.33	2.33	Water play life evolve into important to use.
TP	4.00	4.33	Water in the evolution of life played an important role.
GT	1.33	2.67	Human civilization from the source of the majority of large rivers in the domain.
TP	3.33	4.67	Most of the origin of human civilization in river basin.
GT	2.33	3.00	In human daily life, the water in drinking, cleaning, washing and other side to make use of an indispensable.
TP	3.67	3.33	In human daily life, water for drinking, cleaning, washing and other essential role.
GT	2.00	2.33	Eastern and Western ancient Pak prime material view of both the water regarded as a kind of basic groups into the elements, water is the Chinese ancient five rows of a; the West ancient four elements that also have water.
TP	3.00	3.33	East and West in ancient concept of simple substances regarded water as a basic component elements. Among them, the five elements of water is one of ancient China; Western ancient four elements that also have water.
GT	4.00	4.00	Early cities will generally be in the water side of the establishment, in order to solve irrigation, drinking and sewage problems.
TP	4.67	4.33	Early cities are generally built near the water to solve the irrigation, drinking and sewage problems.
GT	3.0	3.33	Human very early on began to produce a water awareness.
TP	2.67	3.00	Man long ago began to understand the water produced.

Figure 1: Original Google Translate output (GT) together with translations produced by the targeted paraphrase translation process (TP), selected to show a range from strong to weak improvements in fluency.

weight in the translation process, drastically reducing the burden on bilingual translators and lead to a significantly more cost-effective but equally high quality translation process.

6. REFERENCES

[1] J. S. Albrecht, R. Hwa, and G. E. Marai. Correcting automatic translations through collaborations between mt and monolingual target-language users. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–68, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[2] B. B. Bederson, C. Hu, and P. Resnik. Translation by iterative collaboration between monolingual users. In *Graphics Interface (GI) conference*, 2010.

[3] O. Buzek, P. Resnik, and B. B. Bederson. Error driven paraphrase annotation using mechanical turk. In *NAACL 2010 Workshop on Creating Speech and Text Language Data With Amazon's Mechanical Turk*, 2010.

[4] C. Callison-burch, C. Bannard, , and J. Schroeder. Improving statistical translation through editing. In *Workshop of the European Association for Machine Translation*, 2004.

[5] C. Dyer. Noisier channel translation: translation from morphologically complex languages. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, June 2007.

[6] C. Dyer, S. Muresan, and P. Resnik. Generalizing word lattice translation. In *Proceedings of HLT-ACL*, Columbus, OH, 2008.

[7] C. Dyer and P. Resnik. Forest translation. In *NAACL'10*, to appear.

[8] N. Habash and J. Hu. Improving arabic-chinese statistical machine translation using english as pivot language. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[9] P. Koehn. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[10] A.-M. Laurian. Machine translation : What type of post-editing on what type of documents for what type of users. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, 1984.

[11] D. Morita and T. Ishida. Designing protocols for collaborative translation. In *PRIMA '09: Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems*, pages 17–32, Berlin, Heidelberg, 2009. Springer-Verlag.

[12] D. Shahaf and E. Horvitz. Generalized task markets for human and machine computation. In *AAAI 2010*, July 2010.