Position Paper: Improving Translation via Targeted Paraphrasing

Yakov Kronrod
Linguistics and UMIACS
University of Maryland
yakov@umd.edu

Philip Resnik
Linguistics and UMIACS
University of Maryland
resnik@umd.edu

Olivia Buzek
Linguistics and Computer Science
University of Maryland
olivia.buzek@gmail.com

Chang Hu
Computer Science
University of Maryland
changhu@cs.umd.edu

Alex Quinn
Computer Science
University of Maryland
aq@cs.umd.edu

Benjamin B. Bederson
Computer Science and UMIACS
University of Maryland
bederson@cs.umd.edu

Abstract

Targeted paraphrasing is a new approach to the problem of obtaining cost-effective, reasonable quality translation that makes use of simple and inexpensive human computations by monolingual speakers in combination with machine translation. The key insight behind the process is that it is possible to spot likely translation errors with only monolingual knowledge of the target language, and it is possible to generate alternative ways to say the same thing (i.e. paraphrases) with only monolingual knowledge of the source language. Evaluations demonstrate that this approach can yield substantial improvements in translation quality.

1 Introduction

For most of the world's languages, the availability of translation is limited to two possibilities: high quality at high cost, via professional bilingual translators, and low quality at low cost, via machine translation (MT). The spectrum between these two extremes is very poorly populated, and at any point on the spectrum the ready availability of translation is limited to only a small fraction of the world's languages.

There is, of course, a long history of technological assistance to translators, improving cost effectiveness using translation memory (Laurian, 1984; Bowker and Barlow, 2004) or other interactive tools to assist translators (Esteban et al., 2004; Khadivi et al., 2006). And there is a recent and rapidly growing interest in crowdsourcing with non-professional translators, which can be remarkably effective (Munro, 2010). However, all these alternatives face a central availability bottleneck: they require the participation of humans with bilingual expertise.

In this presentation, we discuss a novel use of crowdsourcing that makes it possible to explore the middle ground. We take advantage of a virtually unutilized resource for translation: speakers who know only the source language being translated, or only the target language being translated into, but not both.

The solution we are proposing has the potential to provide a more cost effective approach to translation in scenarios where machine translation would be considered acceptable to use, if only it were generally of high enough quality. There are many real-world scenarios in which a "good-enough" translation really is good enough. These include getting the gist of foreign news, doing first pass translations for wikipedia pages, generating international comments about a product, or rough translations of other content for post-editing by qualified humans.

2 Impact on Translation Professionals

Note that this approach would clearly exclude tasks like translation of medical reports, business contracts, or literary works, where the validation of a qualified bilingual translator is absolutely necessary. Therefore, we do not believe that advances such as this post any threat to the translation profession. In fact, this approach would largely extend the space of possible translation beyond what can be covered by human translators, providing a critical tool for translation of important information into languages

for which translators just do not exist and which are largely commercially ignored.

3 Staking Our Position

The vast majority of technological work is focused on trying to replace the translation process with something fully automatic. This describes virtually all current work within the currently dominant statistical MT paradigm. Yet the weaknesses of that paradigm are well known. Statistical MT breaks the world down into (a) specifying relevant units of translation within a sentence, (b) choosing the right translation units given a set of inputs, and (c) getting those units in the correct order. This leaves aside global properties of the document as well as broader properties having to do with connotations, nuances of usage, and culture.

The trouble with fully human translation, of course, is that it is expensive and requires the availability of appropriate bilinguals. This is overkill for many kinds of translation, as mentioned above, and it fails to take the best advantage of what technology does have to offer. (Translation memories are widely used, but really primitive relative to advances the MT world has made in the last 10 years.)

We propose that most people taking a technological approach to translation are trying to solve the wrong problem. Of course, we do need basic research on fully automatic translation, and there are needs that fully automatic translation can meet now. But the right way to look at MT is as a means to an end. Specifically, a means to obtain translation that meets adequacy criteria of (a) sufficiently high quality, (b) sufficiently low cost, and (c) sufficient availability in the language pairs where it is needed.

With this outlook, the translation problem looks very different. Instead of asking how to improve MT systems, you start by asking how you improve the process of translation for the relevant use cases, a process that can include technology, people – or, most logically, a combination of the two.

Taking that as our position, we quickly arrived at a key observation: for any pair of languages, there are a lot more people who know one or the other, than there are people who know both. And with that supply comes the potential to improve availability and reduce cost, if we can find a way to ensure sufficiently (remember, *sufficiently*) high quality.

What we present in the remainder of the paper is a description of our initial efforts based on that observation. We have succeeded in changing the division of labor so that it is not all technology, nor all human, but rather a balance of the two in which technology does some things that are within its current capabilities, and people do things they can be good at with one language rather than two. At a very high level, the technology provides the cross-language bridge, and people provide the ability to identify errors on the target side and to introduce alternative phrasings on the source side. The results, although small-scale, provide a strong proof of concept that this division of labor can improve on the quality of MT at significantly lower cost.

4 Technology from 1000 feet

We call the method we have developed the targeted paraphrasing translation process, or ParaTrans, for short. In- stead of placing the entire translation burden on a machine (MT) or on a single human (bilingual translators), Para- Trans exploits a new division of labor between machine and human capabilities. The key insight behind the process is that it is possible to spot likely translation errors with only monolingual knowledge of the target language, and it is possible to generate alternative ways to say the same thing (i.e. paraphrases) with only monolingual knowledge of the source language. Operationally, then, translation with targeted paraphrasing includes the following steps.

1. Initial machine translation

   Any automatic translation system can serve in this role. For this paper, we use the Google Translate Research API.

2. Identification of mistranslated spans

   This step identifies parts of the source sentence that lead to ungrammatical, nonsensical, or apparently incorrect translations on the target side.

3. Source paraphrase generation

   This step generates alternative expressions for the source spans identified in the previous step.

4. Generating sentential source paraphrases

   All combinations of source paraphrases from the previous step are multiplied out to provide full sentence paraphrases for each original source sentence.

5. Machine translation of alternative sentences

   The paraphrased sentences are sent through the same MT system, with the best translation hypothesis, according to an agreed upon criteria, is selected.

Notice that with the exception of the initial translation, each remaining step in this pipeline can involve either human participation or fully automatic processing. The targeted paraphrasing framework therefore defines a rich set of intermediate points on the spectrum between fully automatic and fully human translation, of which we explore only a few in this paper.

5 3 Experiments

We conducted three experiments to guage the potential of our approach. We summarize them below

5.1 Pilot Study

We conducted a pilot study with 12 Chinese sentences using the paratrans paradigm. Human participation in this task was accomplished using Amazon Mechanical Turk. The tasks were easy to perform (no more than around 30 seconds to complete on average) and inexpensive (less than $1 for the entire pilot study).

Google Translate (GT) and targeted paraphrasing (TP) outputs were evaluated according to fluency and adequacy of translation on a 5-point scale. The average GT output ratings were 2.36 for fluency and 2.91 for adequacy. Averaging across the TP outputs, these rose to 3.32 and 3.49, respectively. If we consider only the best TP translations then the scores raise to 3.58 and 3.73. Those are respective gains of 1.21 and 0.82 over the baseline initial MT output, a substantial gain.

5.2 Chinese to English Translation Experiment

As a follow up to our pilot study, we conducted an evaluation using Chinese-English test data taken from the NIST MT08 machine translation evaluation. We report on results for 49 sentences that underwent the same targeted paraphrasing process as in the pilot study. The entire cost for the human tasks in this experiment was $5.06, or a bit under $0.11 per sentence on average.

Selecting a single-best paraphrased translation according to Google Translate Research API scores yielded an improvement of 1.68 BLEU points on the 49-sentence test set (TP one-best). A similar oracle-best calculation using TERp for targeted paraphrasing (TP oracle) showed a potential gain of 2.46 BLEU points over the baseline. TERp scores for individual sentences showed improvements for 32 of the 49 test sentences, or 65.3%. For those 32 sentences, the average gain is 8.36 TERp points. Including the sentences with no gain still yielded an improvement of 5.46 TERp points on average.

Our automatic evaluation confirms the subjective ratings results obtained in our pilot study.

5.3 Automatic Error Span Detection

In this experiment, we take a step toward more automated processing, replacing human identification of mistranslated spans with an a fully automatic method. Briefly, we automatically translate source F to target E, then back-translate to produce F in the source language. We compare F and F using TERp and when at least two consecutive edits are found, we flag their smallest containing syntactic constituent as a potential source of translation difficulty.

After identifying the spans and collecting paraphrases from English speakers, we obtained full-sentence paraphrase alternatives for 1000 sentences., which we again evaluated using an Oracle study. We found that better translations were available for 313 out of the 1000 sentences and TP yielded an average TER improvement of 12.16 points, or 3.8 if we include sentences where no gain was obtained. In total, the cost for the human tasks for the study was only $117.48, or a bit under $0.12 per sentence.

6 Conclusion

In this paper we have focused on a relatively less explored space on the spectrum between high qual-

ity and low cost translation, sharing the burden of the translation task among a fully automatic system and monolingual human participants. Our experimental results provide strong support for the argument that targeted paraphrasing can lead to significant improvements in translation. Human judgments also confirmed the availability of better translations. In addition, costs were also kept low, averaging only $0.12 per sentence.

However, we still need to do more work on availability. Mechanical Turk is just one mechanism, and it has various issues, not least of which is people who try to game the system – this is well known among people working on crowdsourcing. We are exploring a variety of options including communities of volunteers. The availability question is closely tied to the question of cost. Again, the crowdsourcing world is facing questions of how much to pay people, in terms of both getting the job done and ethical compensation.

Once you start thinking about the process differently, we can ask ourselves how does one change the technology to take best advantage of it? Right now, we're using the MT system as a black box, but there are clearly avenues one can take to do better. On the MT side, a clear avenue to explore is paraphrase lattices based on human paraphrases (Du et al., 2010; Dyer et al., 2008). On the human side, there is exploring a wider space of contributions that monolingual human participants can make while participating in the process (Bederson et al., 2010).

There are surely other limitations to our approach and ideas that need more exploration. And it is key that they be explored. Yes, a minority of the community has worked for years on human assisted machine translation or machine assisted human translation (e.g. (Bowker and Barlow, 2004; Esteban et al., 2004; Khadivi et al., 2006; Laurian, 1984) and others), but it's time to broaden that conception byeond "assistance". What we need is a high level process, a "system", that involves not just technology plus bilinguals, but technology plus human participants at all levels of linguistic capability. Only then can we begin to reach for the goal, and take the focus off the means.

Acknowledgments

This work has been supported in part by the National Science Foundation under awards BCS0941455 and IIS0838801.

References

Benjamin B. Bederson, Chang Hu, and Philip Resnik. 2010. Translation by iterative collaboration between monolingual users. In *Graphics Interface (GI) conference*.

Lynne Bowker and Michael Barlow. 2004. Bilingual concordancers and translation memories: a comparative evaluation. In *LRTWRT '04: Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training*, pages 70–79, Morristown, NJ, USA. Association for Computational Linguistics.

Jinhua Du, Jie Jiang, and Andy Way. 2010. Facilitating translation using source language paraphrase lattices. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, October. Association for Computational Linguistics.

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

José Esteban, José Lorenzo, Antonio S. Valderrábanos, and Guy Lapalme. 2004. Transtype2 - an innovative computer-assisted translation system. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 94–97, Barcelona, Spain, jul. Association for Computational Linguistics. TT2.

Shahram Khadivi, Richard Zens, and Hermann Ney. 2006. Integration of speech to computer-assisted translation using finite-state automata. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 467–474, Morristown, NJ, USA. Association for Computational Linguistics.

Anne-Marie Laurian. 1984. 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*.

Robert Munro. 2010. Haiti emergency response: the power of crowdsourcing and SMS. Relief 2.0 in Haiti, Stanford, CA.