Beyond bitext: Five open problems in machine translation

Adam Lopez and **Matt Post**
Human Language Technology Center of Excellence
Johns Hopkins University

Abstract

In twenty years, the machine translation (MT) research community has learned a great deal about problems that can be solved with bitext. Yet for many potential MT uses, there is little if any available bitext. In the next twenty years, these uses will become increasingly important, and the research community must marshal its resources to solve the new problems that they present. Specifically, we must assemble large numbers of *small* bitexts for testing systems, rather than small numbers of *large* bitexts for training them. Small bitexts won't solve the new problems alone, but they will help the research community identify the problems that need solving.

1 Is MT a solved problem?

Readers perusing the translation competition results of the annual Workshop on Statistical Machine Translation (WMT) might be tempted to conclude: *MT is a solved problem*. Well-known industrial translation systems, identified in the proceedings as ONLINE-A and ONLINE-B, consistently appear at or near the top of the rankings in every track. Many academic systems fall far behind, and those that beat or tie industrial systems are industry-grade—the winner in three tracks this year used a terabyte language model.[1] Democratization of technology is great, but if the best that academic researchers can do is tie the industrial systems by open-sourcing methods developed in industry, what research problem is the academic community solving? Given their financial incentives and their access to human and computational resources, there is every reason to believe that industry will continue to dominate the question of what to do with big piles of bitext. To the extent that MT reduces to that problem, it *is* solved.

But of course there is more to MT. This raises the question: what problems aren't we likely to solve with a big pile of bitext? Put differently: for what language pairs and domains *don't* we have big piles of bitext? Even if we restrict ourselves to markets with many potential users by focusing only on languages with tens of millions of speakers, there are thousands of possible language pairs. At best, we have substantial quantities of bitext in a few hundred of these. In most of those cases, the bitext is government text. For the vast majority of languages and domains, there is hardly anything.

It may be that scarcity of bitext reflects a lack of interest in translation for some language pairs or domains. However, as Kay (2005) reminds us, two types of people use translation: those who *produce* translations for dissemination in many different languages, and those who *consume* translations in order to understand something in languages other than their own. Most bitext serves the first set of users, consisting mainly of multilingual governments and international corporations. By extension, much of what we do with bitext also serves this set. Kay remarks that "*what the very word translation means for these two sets of people is entirely different. And I just would like to hope that you, the computational linguists of the future, will keep in mind the needs of both of these very worthy communities.*" In short, if we are to serve the needs of more than politicians and marketers, we must look beyond large bitexts.

These issues are not new, but to what extent is the community engaged in them? As empiricists, we prefer to ground the answer to this question in data. Following Bender (2009), we surveyed long and short papers at ACL 2013, identifying 51 with MT experiments. We report below the number of papers with particular test set characteristics, by language and domain, for all cases occurring more than once.[2] We found individual papers with experiments on Czech, Farsi, Finnish, Italian, Kazakh, Kirghiz, Korean, Persian, Russian, Turkish, and Uyghur–though only three papers accounted for most of these. A few

[1] Notably, this winner was once a Google engineer.

[2] Some study multiple languages, so numbers do not sum to 51.

Language pair	government/ news	other
Chinese to English	25	3
Arabic to English	5	3
Spanish to English	6	-
German to English	5	-
French to English	5	2
Japanese to English*	4	-
English to German	3	-
Czech to English	3	-
Urdu to English	2	-

* Japanese to English experiments were on patents.

studied translation out of English, or between non-English languages. However, most relied on large bitexts. If the research community is truly engaged on problems of scarce bitext, its premier research venue does not reflect this.

2 Five Open Problems in MT

With so many languages and domains underrepresented in bitext, there is a wealth of possible research problems to address. Several well-known problems are, fundamentally, problems of scarce bitext.

Translation of low-resource language pairs. The most straightforward example of scarce bitext covers most of the world's language pairs.

Translation across domains. Translation systems are not robust across different types of data, performing poorly on text whose underlying properties differ from those of the system's training data.

Translation of informal text. People want to read blogs, social media, forums, review sites, and other informal content in other languages for the same reasons they read them in their own: the motivations are many. However, informal bitexts are scarce.

Translation into morphologically rich languages. Most MT systems will not generate word forms that they have not observed, a problem that pervades languages like Finnish, Arabic, and German.

Translation of speech. Much of human communication is oral. Even ignoring speech recognition errors, the substance and quality of oral communication differs greatly from that found in most bitext.

3 A Challenge for the MT Community

One solution to the open problems above is to develop large bitexts whenever we encounter them. This returns us to the status quo, a setting where the academic research community is at a disadvantage against industry resources, where brilliant engineering prevails over models and methodology and eclipses the goal of increased scientific understanding. Moreover, this solution necessarily ignores the long tail of language pairs and domains in favor of a privileged few. We simply cannot develop large bitexts for every language pair and domain.

However, work on crowdsourcing has shown that it is quite feasible to develop *small* bitexts. We advocate the development of many small, focused bitexts to be used as test sets in languages and domains where large bitexts do not exist. Such efforts have appeared here and there. Recent examples include datasets produced by the 2012 workshop on Domain Adaptation for MT (Braune et al., 2012) and the WMT 2011 featured translation task of Haitian Creole SMS messages (Callison-Burch et al., 2011). But these efforts must be sustained and primary if they are to have any long-term effects. The problems we find in these test sets will be hard to solve, since we will have no large bitexts to fall back on. They will likely require creative modeling of language data other than bitext—a direction suggested by Brown et al. (1993) that that has been overshadowed by the focus on large bitexts. This might reveal difficult modeling and learning problems in the effort to exploit data of a type quite different from that found in the prediction problem we want to solve. A community-wide focus on such difficult problems is risky, but the rewards for both MT researchers and users could be immense.

Acknowledgements

Thanks to Pushpendre Rastogi for comments on a previous draft of this opinion piece.

References

E. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proc. Interaction between Linguistics and Computational Linguistics*.

F. Braune, M. Carpuat, A. Clifton, H. Daumé III, A. Fraser, K. Henry, A. Irvine, J. Jagarlamudi, J. Morgan, C. Quirk, M. Razmara, R. Rudinger, A. Tamchyna, and G. Foster. 2012. Domain adaptation in

statistical machine translation. CLSP Workshop Report.

P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, M. J. Goldsmith, J. Hajic, R. L. Mercer, and S. Mohanty. 1993. But dictionaries are data too. In *Proc. of HLT*.

C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proc. of WMT*.

M. Kay. 2005. A life of language. *Computational Linguistics*, 31(4):425–438.