Return of the Linguist: Toward scalable parallel corpus generation for machine translation

Anonymous EMNLP submission

Abstract

The large-scale generation of synthetic parallel corpora for data-driven machine translation is proposed. A rudimentary system for doing this is presented, with mechanisms for syntactic complexity in the spirit of classical rule-based machine translation and high-quality sentence pairs in the spirit of example-based machine translation. Preliminary experiments integrating these synthetic corpora into a shared task are reported. Promising future directions are discussed, in particular, large-scale contributions from "amateur linguists" via crowdsourcing.

1 Motivation

As early as Jelinek's (in)famous quote during the 1980's that "*Every time I fire a linguist, my performance goes up*," both linguists and linguistics have been increasingly marginalized in natural language processing (NLP). More recently, for example, Collobert et al. (2011) mention the possibility "*[ideally]...to learn from letter sequences rather than words*," illustrating an AI-motivated desire to move away from even the most fundamental of linguistic concepts. A specific instance of this overall trend is the gradual shift of machine translation (MT) away from classical linguistic rule-based approaches to contemporary machine learning approaches (Hutchins, 2005), which underlie statistical phrase-based systems like Moses (Koehn et al., 2007) and more recent neural network systems (Sutskever et al., 2014; Cho et al., 2014). But in the end, even the most cutting-edge machine learning MT system relies crucially on multilingual parallel corpora, typically taken from external sources like multilingual parliamentary proceedings. This results in what we perceive to be a disproportionate amount of effort being dedicated to the machine learning component of machine translation, with input corpora being treated almost as an afterthought.[1]

One of the major advantages of data-driven MT is that it removes the need for linguists to develop a new set of rules for an MT system to support a new language. In a sense, though, the rule-based MT linguist's job has really been "outsourced" to third-party translators, who implicitly encode linguistic information in parallel corpora. We believe there is still a significant "in-house" role for linguists to play in corpus-based machine translation, and that is to direct the crafting of high-quality, "designer" synthetic parallel corpora, leveraging technologies like crowdsourcing to maximize throughput. Such a division of labor would let linguists be linguists, with MT system development proceeding in parallel.

Large, high-quality synthetic parallel corpora could be useful as:

- Training sets for corpus-based MT systems, particularly for low-resource language pairs
- Benchmark tasks for MT evaluation, in the spirit of Facebook's bAbI project (Weston et al., 2015)

[1] Of course, these parallel corpora represent hundreds of translator-hours of work. However, such translations are optimized for human consumption, not machine learning, with no regard for such niceties as sentence alignment.

- Translation memories for computer-assisted translation
- "Language archives": New NLP methods are developed every few years; linguistic theories change every decade or so; parallel corpora live forever (like the Rosetta Stone).

We believe that natural language processing in general and machine translation in particular are particularly amenable to artifical data generation. Unlike many machine learning applications (image recognition, finance), NLP data sets are manmade to begin with; humans themselves *are* the primary natural producers and consumers of raw written language data. Translation is further privileged as perhaps the only NLP task for which millions of people (foreign-language learners) already undergo years of formal training, which has implications for scalability via crowdsourcing (Section 5).

Other practical considerations shine a favorable light on parallel corpus generation. As the "inverse" of translation, multilingual generation is also an easier task in some senses. For example, lexical ambiguity can be largely alleviated,[2] segmentation of languages like Chinese (Tseng et al., 2005; Chang et al., 2008) becomes a non-issue, and there is the potential for generating annotated corpora with zero annotation error. Parallel corpora also provide a natural avenue to interface contemporary MT systems with the countless linguist-years of expertise that have been invested into rule-based MT systems. And unlike the various different MT methods, which are by nature competitive, corpora are inherently collaborative.

This paper is organized as follows. Section 2 discusses previous related work, Section 3 describes our rudimentary parallel corpus generator, Section 4 gives preliminary results, and Section 5 proposes future work.

2 Related Work

The idea of using synthetic parallel corpora in machine translation is hardly new. Previous attempts at creating synthetic parallel corpora have used rule-based MT (RBMT) systems (Hu et al., 2007; Dugast et al., 2008; Murakami et al., 2009; Rubino et al., 2014), but the input to the RBMT systems was again dictated by external corpora in the desire to produce immediate results. A longer-term approach that may pay dividends (Section 1) would be to run RBMT systems on specially chosen sentences for which they are known to produce high-quality translations, or perhaps eventually directly encoding their rules into carefully crafted parallel corpora.

Our parallel corpus generation system (Section 3) is very similar to the recursive equivalence class approach to example-based MT (Carl and Way, 2003) used by Brown (1999). Our generator has more of a rule-based flavor, using a somewhat more varied grammar that also admits the use of participial noun modifiers. But more importantly, we believe this line of work was ahead of its time and consider our main contribution to be bringing attention to its potential value to contemporary MT.

Equivalence classing is in fact an integral part of the standard pipeline of Moses (Koehn et al., 2007), the open-source statistical MT system. During training, the `mkcls` component adds "*an 'example-based' touch to the statistical approach*" by clustering words into automatically-generated equivalence classes to ameliorate data sparsity issues (Och, 1999). However, these clusters are trained only at the word level and not for phrases, and the clusters themselves are typically only used during alignment. Furthermore, since they are learned in an unsupervised manner from the parallel corpus, they can still benefit from the addition of more high-quality training examples, synthetic or otherwise.

In terms of freely available multilingual natural language generators, the most well-developed one we found was KPML (Bateman, 1997). But crafting sentences using KPML is a lengthy process, involving a well-developed linguistic theory to explain the surface forms. In contrast, our primary goal is to generate large parallel corpora of surface forms quickly.

In the next section, we describe our own primitive attempt at building a multilingual generator from scratch. While we try to model straightforward linguistic phenomena compositionally, we also adopt the pragmatic attitude, in the tradition of example-based MT (Carl and Way, 2003), that sometimes the most effective way to represent a translation phenomenon is simply to write it down. Thus, with

[2] Ideally the generator would "know" *a priori* whether it is writing a sentence about, say, (financial) banks or (river) banks.

the long-term goal of scalability, we attempt to construct a system sufficiently flexible to write some sentence pairs in large quantities through syntactic variations and others with “hand-crafted” quality through example-based templates.

3 A Nascent Multilingual Generator

As an example of how our synthetic parallel corpus generation system works, consider the following English-Chinese template pair:

S V O .
S 把 O V 。

Our system then expands the subject S and object O as noun phrases, adding adjectival and participial modifiers as desired. Semantic constraints can be imposed from the verb V as well as externally. The generator can then vary the template lexically[3] and syntactically, allowing a large number of sentence pairs to be generated from a single template. We can also trade quantity for quality by adding more literals to the template, in the spirit of example-based MT.

We adopt an extremely simplistic view of language as nouns, verbs, and their modifications and elaborations. The fact that verb phrases can also serve as nouns (gerundives) and modifiers (participles) appears to be a major source of syntactic complexity, and we made it a priority to implement participial modifiers. To capture recursive structures like participles, we adopt a rudimentary tree representation that is something intermediate between constituency and dependency trees, with constituent-like modifiers that also maintain dependency-like links to their targets. We plan on eventually open-sourcing both our generator code and its accompanying data.

4 Preliminary Experiments

We now consider training set amplification as a first application, using the Chinese to English MT task of IWSLT 2015 (Cettolo et al., 2015). This task uses the WIT^3 corpus (Cettolo et al., 2012), which consists of multilingual subtitles of TED talks covering a wide variety of subjects. The freely available MultiUN corpus (Eisele and Chen, 2010) from the OPUS project (Tiedemann, 2012) was also used for comparison.

To investigate the relative effect of our synthetic parallel corpora, we chose a simple baseline system and fixed all other variables. Namely, we used the baseline configuration of Moses (Koehn et al., 2007) as described in (Cettolo et al., 2015), with the exception of changing the target language model to KenLM (Heafield et al., 2013), which we found to be faster and stabler at decoding time than IRSTLM (Federico et al., 2008). For the remaining unspecified model settings, we used the `grow-diag` alignment heuristic recommended by Chang et al. (2008) and the `mslr-fe` configuration for lexicalized reordering. Chinese text was preprocessed using the latest version of the Stanford Segmenter (Tseng et al., 2005; Chang et al., 2008). Phrase and lexicalized reordering tables were interpolated using the `tmcombine` script included with Moses (Sennrich, 2012), with the interpolation tuned on the `dev2010` data set. BLEU scores were computed as described by Cettolo et al. (2015) using the IWSLT 2015 progress data set (`tst2014`).[4]

	BLEU		BLEU
25% corpus	9.95	Official	11.43
+ 75%	**11.47**	Our baseline	11.54
+ MultiUN	10.76	+ MultiUN	**11.67**
+ Generated	9.91	+ Generated	11.46

Table 1: (Left) Chinese-to-English BLEU scores for our baseline Moses configuration with phrase and reordering tables trained on a random 25% subset of the IWSLT15 training corpus and interpolated (+) with tables for the remaining 75% (˜150K) of the in-domain corpus, the MultiUN corpus (a ˜1M line subset), and a synthetic corpus from our generator (˜5M lines). (Right) Same as the left, but starting from tables trained on the entire IWSLT15 corpus. Official published scores of the IWSLT15 baseline system are also given for comparison. Qualitatively similar trends for NIST and TER scores were also observed (omitted for clarity).

The results of these experiments are shown in Table 1. For the table on the left, note that adding in-domain data improves performance significantly more than a much larger amount of out-of-domain

[3] Rewrite probabilities were set by hand; a more principled approach might use, say, a generative PCFG parser.

[4] We chose the `tst2014` test set to facilitate comparison with published results from both IWSLT 2014 and 2015 in future work.

data (MultiUN), a commonly reported phenomenon (Sennrich, 2012). TED talks, which are prepared lectures, fall somewhere between the conversational language of movie subtitles and the formal written language of UN resolutions. As such, the OpenSubtitles[5] 2016 corpus (Lison and Tiedemann, 2016) from the OPUS project was also out-of-domain; results for this corpus were similar to MultiUN and are not shown. For the table on the right, note the diminishing returns of out-of-domain data as the amount of in-domain data increases. These observations underscore the importance of domain overlap and hint at the potential value to be found in the amplification of in-domain corpora.

Our generated corpora do not yet make a significant impact on the final BLEU score,[6] indicating that our parallel corpus generator has a long way to go. Development on our generator has so far focused on "rule-based" syntactic diversity and growing a lexical database tailored to the training set. Given the resulting scores, a change of direction seems in order; work on harvesting more "example-based" templates from the IWSLT15 training corpus is currently in progress.

5 Proposed Future Work

Many linguistic constructions have yet to be implemented in our rudimentary/primitive generator (for example, past tense is only partially supported). Future development may be expedited by adapting work from existing rule-based MT or natural language generation systems. Dynamic, on-the-fly corpus generation may also have a role to play in online or curriculum learning settings.

Deep learning methods like RNN encoder-decoders (Sutskever et al., 2014; Cho et al., 2014) are but the latest in a succession of corpus-based approaches to MT, following in the footsteps of example-based and statistical machine translation; in this era of Big Data it is unlikely that they will be the last. But all such methods share a reliance on parallel corpora, and we believe parallel corpus generation will remain a relevant pursuit as even newer methods emerge.

As alluded to in Section 1, parallel corpora have the innate potential for open collaboration; not only can they be expert-sourced to linguists, but they can also crowdsourced to "amateur linguists." One particularly promising application domain is conversational language, for which syntactic and lexical diversity is expected to be relatively low compared with more formal domains. Multilingual lexical databases for slang terms, in particular, would be particularly suitable for crowdsourcing. At the very least, crowdsourcing should be strongly considered for post-editing small corpora used as potentially noise-sensitive components of shared tasks, like development or test sets. For shared evaluation tasks with many language pairs, crowdsourcing may also be the only practical way to do this.

We believe the most promising path to large-scale parallel corpus creation is through educational crowdsourcing. As with the success story of reCAPTCHA (von Ahn et al., 2008), we believe this will scale best when there is a natural harmony between data supply and demand. Unlike other staple NLP tasks like parsing or part-of-speech tagging, MT is particularly well-suited to this, with English being taught as a foreign language in classrooms around the world. The data format (English itself) is widely agreed upon, and there is even an existing market, a vast untapped resource of student problem sets that is yet to be harnessed by machine learning.

Large quantities of supervised data underlie state-of-the-art performance in many NLP tasks. While current paid approaches like Amazon's Mechanical Turk are already a viable way of obtaining crowd-sourced data, we believe it would be well worth the initial investment of time and resources to build a educational app compelling enough to become widely adopted. This could then serve as a platform from which to harvest supervised natural language data from its natural source. Compared with current paid approaches, the cost per datum could be significantly reduced or perhaps even become negative (a profitable app), although the value of the harvested data itself might be worth subsidizing. We consider the construction of such a potentially game-changing source of supervised NLP data to be an open challenge to the NLP community in general.

[5] http://www.opensubtitles.org/

[6] The sub-baseline scores shown are for our most recently generated synthetic corpus; scores slightly higher than our baseline have also been obtained on older generated corpora. Both of these outcomes seem to be within experimental error bars.

References

John A. Bateman. 1997. Enabling technology for multilingual natural language generation: The kpml development environment. *Nat. Lang. Eng.*, 3(1):15–55, March.

Ralf D. Brown. 1999. Adding linguistic knowledge to a lexical example-based translation system. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22–32.

Michael Carl and Andy Way, editors, 2003. *Recent Advances in Example-Based Machine Translation*. Springer Netherlands, Dordrecht.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 2–14, Da Nang, Vietnam, December.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2008. Can we relearn an rbmt system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 175–178, June.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pages 2868–2872.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1618–1621, Brisbane, Australia, September.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Xiaoguang Hu, Haifeng Wang, and Hua Wu. 2007. Using rbmt systems to produce bilingual corpus for smt. In *Proc. of EMNLP-CoNLL 2007*, pages 287–295.

John Hutchins. 2005. The history of machine translation in a nutshell.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris 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, June. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Jin'ichi Murakami, Masato Tokuhisa, and Satoru Ikehara. 2009. Statistical machine translation adding pattern-based machine translation in chinese-english translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2009)*, pages 107–112.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 71–76, Bergen, Norway, June.

Raphael Rubino, Antonio Toral, Nikola Ljubešić, and Gema Ramírez-Sánchez. 2014. Quality estimation for synthetic parallel data generation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference* *of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171, Jeju Island, Korea, October. Association for Computational Linguistics.

Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698.