German-to-English Translation

A Driving Force Behind Machine Translation

Kenton Murray

Machine Translation (MT) is a subfield of computer science that studies automatic translation from one human language to another (or between human languages and computer languages). It is a broad interdisciplinary field that relies not only on core computer science areas such as coding and algorithm design, but also on broader academic disciplines such as statistics and linguistics. Machine Translation can be done with any pair of languages, and yet German-English translation is the most studied combination in the field. The economic demand for translation between these languages is not significantly higher than for other frequently translated languages, like Spanish, French, Chinese, and Arabic. In this paper, I argue that the overrepresentation of this language-pair is the result of three core phenomena: resources, people involved, and linguistic interest. No single phenomenon is solely responsible for this effect, but the unique interplay of the above three factors, has made German-English translation one of the driving forces behind new developments in Machine Translation.

Developing new MT tools requires the analysis of thousands and thousands of examples, and therefore finding numerous high-quality translations and sources is the starting point for most MT research. The ideal resource is an online collection of bilingual texts, known as »bitexts.« The Europarl Corpus is one of the best funded large-scale translation projects available to the public online, and of its 21 languages, German-English are the most studied. The reasons for this include the number of German-speaking researchers and the widespread economic and political interest in English. The difference in the word order, among other linguistic differences, makes it an interesting challenge for MT research.

The predominance of German-English translation in the Machine Translation literature is staggering. Even though there are more than 5 000 languages on Earth, very few are ever used in MT, and none as much as German-English. One of the most important annual meetings of MT researchers is the Conference on Machine Translation, and German-English research is featured prominently there.[1] At the 2018 conference, 67 % of the peer-reviewed research papers

1 | Ondřey Bojar et al.: Proceedings of the Third Conference on Machine Translation. Association for Computational Linguistics, October 2018, Brussels (B).

at least discuss German translation, while 48 % actually show results of their method on a German translation task. Ironically, over 20 % of the papers that do not mention German at all are from German universities. This is common in other conferences and publication venues as well, anchoring German-English as an important influence on advances within the field.

Resources

Though methods within the field vary substantially, the vast majority of MT systems generally rely on probabilistic models as opposed to hand-engineered grammar rules. For instance, »7:00 pm works for me.« would best be translated »19 Uhr passt mir.« and »19 Uhr klappt für mich«. Though a human translator may consider both to be correct, native speakers might rate »19 Uhr passt mir.« more idiomatic. A well-trained MT system would hopefully replicate this behavior and assign both high probabilities, but give a higher probability to former over the latter. Ideally, it would score an incorrect translation such as »19 Uhr ist vorbei.« as very unlikely with a score near 0. These probabilistic models are a generalization of a wide class of MT models, consisting of both traditional systems as well as the recently popularized methods based off of *Artificial Intelligence* (AI) and Neural Networks. At their core, they all still rely on decades old methods of probabilistically scoring for potential translations. All of these models learn these probabilities by being exposed to millions of translated sentences through the corpus resources we call »bitexts«.

Bitexts are very large, sentence-aligned, parallel corpora generated by expert human translators. They serve as the data for which MT algorithms and models are both trained and evaluated on. Due to the complexity of language, to get enough training examples to learn meaningful probabilities, MT systems generally rely on millions of translated sentences. A low-resource bitext will still generally have 200,000 translated sentences and MT systems trained on this will only have mediocre performance. To put that number in context, 200 000 sentences is roughly the entirety of Shakespeare's works five times over. Naturally, creating adequately large bitexts is very expensive. In 2001, Ulrich Germann estimated that the cost of creating a 200 000 sentence bitext was 1.5 Million USD.[2]

As in any academic field, the availability of resources, both financial and data, is a major factor in how a field advances. Within machine translation, much of the funding comes from North American and European institutions,

2 | Ulrich Germann: Building a statistical machine translation system from scratch: how much bang for the buck can we expect? In: Proceedings of the workshop on Data-driven methods in machine translation-Volume 14. Association for Computational Linguistics 2001; Tomer Levinboim: Invertibility and Transitivity in Low-resource Machine Translation. Diss. University of Notre Dame 2017.

which are likely to have interest in German. But most importantly, the availability of large German bitexts drives the field. The most common source of German bitexts comes from the European Union parliament, which translates all proceedings into other EU languages. By virtue of these being government records, transcripts are freely available on the European Union's website. MT researchers have created the »Europarl Corpus« by collecting years of these transcripts using a web spider – a program that starts on a webpage and automatically navigates links.[3] Importantly, the data is open and free to use thanks to the permissibility of government documents, which makes it readily accessible to all researchers regardless of institutional affiliation or financial support. The existence of the Europarl corpus is one of the main reasons the field of MT has advanced as much as it has, and naturally it contains other languages as well. However, German and French are the two most commonly used in the MT literature. The existence of this large dataset has therefore precipitated many advancements, yet its existence alone cannot solely explain the impact the German-English language pair on the field of MT (since, after all, the French-English corpus is comparable in size).

People Involved

A crucial way that MT researchers evaluate proposed and published methods in the field is through the use of a »shared task«. A shared task consists of an agreed upon set of training data (usually a bitext plus other linguistic resources). Research teams spend weeks or months designing and training systems for the agreed upon language pair – often including recent advances in the field as well as the latest published results. Then, on a certain date, every team is given some new, unseen, monolingual data and their system must translate it. All of the translations are compared against each other and scored using a variety of metrics. Shared tasks are great from a scientific perspective as they serve as an apples-to-apples comparison. In other words, all experimental settings and parameters are identical so it allows the field to make stronger claims about which methods are better.

Furthermore, with the release of a new test, shared tasks prevent stagnation (»overfitting«) on a single dataset. Though there are a few every year, organized by academics and funding agencies alike, the longest running and most popular is the WMT Translation Task.[4] In part, this is because evaluation is actually done by humans looking at system outputs and ranking them, as opposed to automatic metrics that look at overlaps between the system outputs and a sin-

3 | Philipp Koehn: Europarl: A parallel corpus for statistical machine translation. In: MT summit 5 (2005).

4 | Ondřey et al., Proceedings of the Third Conference.

gle, immutable translation without human intervention. Creating a test set for a shared task, as well as manually evaluating systems, is a time-consuming and labor-intensive task. 2018 saw the 13th annual shared translation task of news. As in other years, there were multiple languages in the shared task, including German-English. Even so, German-English translation is the only language pair that has been featured in all 13 years of the shared task. In part, this consistency is another reason why proposed methods and published papers use German translation even when they are not competing in the task. Running a shared task is a lot of work. As any academic can attest, professional service is often undervalued. A major reason that German has been included for so long is due to the work that many German Computer Scientists have put in. This is the second reason that German-English is so important to MT – the professional service of German speaking academics.

Linguistic Features

Finally, linguistic interest also motivates the use of German for translations. Most of the published literature in the field of Machine Translation is in English. Translating between English and German allows researchers to observe how proposed methods and systems are able to handle various linguistic phenomenon. For instance, German exhibits more inflectional variety than English, so we can investigate whether a system has learned the proper declension of definite articles or a noun's gender and case. Reordering of words in a translation is also something that is difficult for many MT systems, so the behavior of a method on non-finite verbs in German can be interesting. Other aspects such as compound nouns have motivated researchers to focus on splitting words, which has helped research across other languages.[5] Finally, the fact that both English and German use similar writing systems facilitates even non-German speakers to inspect system outputs and data – which is an added layer of complexity when working in bitexts that don't, such as Arabic or Chinese. Though the availability of resources, as well as the work of German speaking academics are key to German's impact on MT, the fact that it linguistically motivates research questions is also integral to its prevalence.

Overall, it is the interplay of all three factors that has caused German-English to be a significant driving force behind advances in Machine Translation. This unique combination of available resources, people, and linguistics has created a ripe environment for researchers to focus on methodological improvements. However, though the field of MT has made significant advances in recent years,

5 | Rico Sennrich/Barry Haddow/Alexandra Birch: Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: *Long Papers*) 1 (2016).

systems are not at parity with human translators and there remains a lot of open research questions.[6] Assuming that there continue to be German bitexts created, as well as motivated and devoted German speaking researchers willing to devote their time to professional service, German-English will continue to be a driving force behind the improvement of Machine Translation capabilities.

6 | Antonio Toral et al.: Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers 2018.