Multilingual Subtitling in the Age of Google Translate

Martin Volk (University of Zurich, Switzerland), Arantza del Pozo and Rodrigo Agerri (Vicomtech Research Center, Spain)

The last two decades have seen a revolution in building Machine Translation systems, that is computer systems that automatically translate from one language to another. The revolutionary approach is called Statistical Machine Translation and allows for the fully automatic construction of Machine Translation systems when given large amounts of human-translated text. Google Translate is the most prominent example of this new approach. Previously, the construction of a Machine Translation system required a large bilingual dictionary and elaborate grammar rules for analyzing the input sentence and for transfer to the target language. Both are costly and time-consuming to create.

The new statistical approach automatically "learns" translation correspondences from human-translation examples. For instance, given a large collection of English texts plus German translations, the computer system will first establish cross-language correspondences between English and German sentences which is similar to the alignments known from translation memory systems. Subsequently, the system computes word correspondences (which word was translated with which word or phrase). These word alignments are the basis for chunking the parallel sentences into word sequences. All this is done in the preparation phase of the Machine Translation system. During the actual translation phase the input sentence is also cut into chunks, and the corresponding target language chunks are reassembled and, if necessary, reordered to suggest translation hypotheses. Finally, a statistical target language model helps in ranking these translation alternatives in order to determine the best translation.

For example, when we have the English – German subtitle pairs "*Maybe I'll meet him tomorrow. – Vielleicht werde ich ihn morgen treffen*" and "*Chris wants to start up a band – Chris möchte eine Band gründen*", we have the necessary pieces to translate "*Maybe I'll start up a band*" correctly as "*Vielleicht werde ich eine Band gründen*". Note that the difference in word order and the idiomatic correct translation of "*start up*" in this context are resolved by using the chunk pair "*start up a band – eine Band gründen*".

Machine Translation works with varying degrees of success on different text types. As it turns out Statistical Machine Translation is well-suited for translating subtitles, for various reasons. First, subtitles are relatively short textual units, much shorter than average sentences in technical documents or newspapers. Second, subtitles can easily be aligned because of the time codes. Finally, subtitles are surprisingly repetitive. In a collection of 1 million English TV subtitles we found that 10% of the subtitles occur more than once.

This raises the question whether one should use Google Translate to translate subtitles. Google Translate is attractive since it is a free web-service for a large number of languages (57 languages at the time of writing). For this matter and because of its speed and ease of use it has arguably become the world’s best known translation system in recent years. Moreover, Google has scored amongst the top systems in Machine Translation competitions for language pairs as diverse as Chinese – English and Arabic – English. But Google Translate is and remains a general-purpose Machine Translation system that cannot match the quality of special-purpose translation systems for subtitles.

Google collects all kinds of translated documents from the web and analyses them as the basis for their Statistical Machine Translation systems. For some language pairs they probably even use an intermediate language when building the necessary bilingual language resources. However, it is well known that Statistical Machine Translation systems work best for the textual domain that they are trained on (Koehn 2010). The input texts largely determine the output texts. After all, Statistical Machine Translation is a recycling approach. In that respect it builds on the same idea as translation memories. But while translation memory systems can only retrieve translation units as a whole, Statistical Machine Translation reuses word sequences of arbitrary lengths and reassembles them into new translations.

Special-purpose Machine Translation systems that are built on high-quality human-translated subtitles result in a higher Machine Translation quality than general Machine Translation systems. Experience tells us that 1 million subtitles (about 10 million words) of translated text is an ideal starting point for building a Statistical Machine Translation system (see (Volk 2008)). If only smaller amounts are available, this can be partially compensated if other resources, e.g. other parallel corpora or bilingual word lists, are available. Building a Statistical Machine Translation system for subtitles on smaller amounts might be worthwhile when we consider that it is a machine learning approach. The performance of such self-learning translation systems improves as new human-translated subtitles come in.

We have built special-purpose translation systems for Scandinavian languages (Swedish, Norwegian, and Danish, (Volk 2008)) and also for English to Swedish translation. We found that these systems lead to productivity increases of around 25% for the subtitle translators. We are aware that post-editing Machine Translation output changes the translators’ working conditions considerably. Checking and correcting machine output is often seen as restricting the translator’s creativity and freedom. On the other hand, the machine frees the translator from repetitive tasks such as translating the same subtitle over and again. For instance, in 1 million subtitles we found “Are you okay?” 102 times, “What are you saying?” 39 times and “It wasn’t your fault” 10 times.

We observe a trend that commercial providers combine translation memory systems with Machine Translation systems. When the translation of the complete unit cannot be found in the memory and the fuzzy matching score drops below a certain threshold, then Machine Translation takes over. A number of initiatives (like tausdata.org) and companies offer web-based translation memory services. This facilitates the exchange of translation memories and the access to large collections. However, issues of quality control must be resolved, and ultimately trust need to be established over time.

We do not expect leaps in the output quality of Machine Translation systems in the foreseeable future. Therefore it does not make sense to wait until “next year’s” system will fulfill all wishes.

Investing in Machine Translation now will enable the full reuse of the wealth of previous translations.

There is a large body of research on Machine Translation. For example, there are interesting developments in combining the statistical approach with more linguistic knowledge. There are others that work on more practical issues like automatic confidence estimates that would allow the computer to deliver only good translations and not to translate “difficult” sentences by suppressing presumably bad translations that fall below a certain threshold. This functionality will be similar to the fuzzy match scores in translation memory systems that are popular to filter only for the best hits.

There is no doubt that Google Translate is the most visible sign of a new era of Machine Translation. MT technology is omnipresent and used by people from all ages and backgrounds, from school children to managers. Free and easy access to automatic translation for a large number of language pairs has opened the way for new applications such as instant translations of web pages or the use of automatic translation in language teaching.

Another interesting development is the combination of machine translation with crowd translation (e.g. dotSub, Speaklike), the distribution of translation tasks to many translators throughout the web. If we can establish a methodology for finding the best translations out of the crowd, then this might well be the route to the future and influence the translation business significantly.

Machine Translation is on our doorstep, while other automation methods are further away. Automatic speech recognition for auto-captioning (as offered by YouTube) is in a preliminary state and so is the automatic conversion of transcripts into subtitles. The latter requires shortening and perhaps re-phrasing the transcript while maintaining the original meaning.

Conclusion

We have introduced Statistical Machine Translation as a new and revolutionary approach to machine translation, and we have argued why it is well suited for subtitles. In times of pressure on subtitling prices, the integration of Machine Translation is an opportunity to increase translator productivity or – from a management perspective – to reduce subtitle translation costs. Systems that reuse and reassemble large amounts of human-translated subtitles can be built quickly and be profitably integrated in the subtitle translation workflow. Because the performance of Statistical Machine Translation systems is largely dependent on the discourse domain of the training data, this permits the development of special translation systems of subtitles.

References

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press.

Volk, Martin. 2008. The Automatic Translation of Film Subtitles. A Machine Translation Success Story? In: *Resourceful Language Technology*: *Festschrift in Honor of Anna Sågvall Hein.* Uppsala. (https://www.zora.uzh.ch/8817/)