On Statistical Machine Translation and Translation Theory

Christian Hardmeier
Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden
first.last@lingfil.uu.se

Abstract

The translation process in statistical machine translation (SMT) is shaped by technical constraints and engineering considerations. SMT explicitly models translation as search for a target-language equivalent of the input text. This perspective on translation had wide currency in mid-20th century translation studies, but has since been superseded by approaches arguing for a more complex relation between source and target text. In this paper, we show how traditional assumptions of translational equivalence are embodied in SMT through the concepts of *word alignment* and *domain* and discuss some limitations arising from the word-level/corpus-level dichotomy inherent in these concepts.

1 Introduction

The methods used in present-day statistical machine translation (SMT) have their foundation in specific assumptions about the nature of the translation process. These assumptions are seldom discussed or even made explicit in the SMT literature, but they have a strong influence on the way SMT models implement translation. This paper studies the relation between current approaches to SMT and major developments in translation studies. We begin with a brief overview of the most important milestones in translation theory and show that the concept of *word alignment* embodies a view of translation that is strongly related to notions of *translational equivalence* popular among translation theorists of the 1960s and 1970s. Defined in terms of an equivalence relation, translation is seen as an essentially "transparent" operation that recodes a text in a different linguistic representation without adding anything of its own, a view that ignores much of the complexity of the decision making processes involved in translation. We show how SMT works around this problem by using the concept of *domain* as a corpus-level catch-all variable and discuss why this approximation may not always be sufficient.

2 Perspectives on Translation

It has been recognised since antiquity that word-by-word translation is generally inadequate and that a higher level of understanding is necessary to translate a text adequately into another language. The fourth century church father and bible translator Jerome made a conceptual distinction between translating "word for word" and "sense for sense" (Jerome, 1979), which remained fundamental for theoretical discussions of translation until the first half of the 20th century (Bassnett, 2011).

Until the 1990s, translation was seen as an act of *transcoding* ("Umkodierung"), whereby elements of one linguistic sign vocabulary are substituted with signs of another linguistic sign vocabulary (Koller, 1972, 69–70). The principal constraint in this substitution is the concept of *equivalence* between the source language (SL) input and the TL output:

> Translating consists in reproducing in the receptor language the closest natural equivalent of the SL message, first in terms of meaning and secondly in terms of style. (Nida and Taber, 1969, 12)

Nida and Taber (1969, 12) emphasise that the primary aim of translation must be "reproducing the message", not the words of the source text. According to them, translators "must strive for equivalence rather than identity" (Nida and Taber, 1969, 12). They stress the importance of *dynamic equivalence*, a concept of functional rather than formal equivalence that is "defined in terms of the degree to which the receptors of the message in the receptor language respond to it in substantially the same manner as the receptors in the source language" (Nida and Taber, 1969, 24). Koller (1972)

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 168–172,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

adopts a similar position. Instead of highlighting the message of the source text, he focuses on *understandability* and defines translation as the act of making the target text receptor understand the source text (Koller, 1972, 67).

The end of the last century brought about an important change of viewpoint in translation studies, which has been named the *cultural turn* (Lefevere and Bassnett, 1995; Snell-Hornby, 2010). Equivalence as a purely linguistic concept was criticised as deeply problematic because it fails to recognise the contextual parameters of the act of translating; it was called an "illusion" by Snell-Hornby (1995, 80), who also pointed out that the formal concept of equivalence "proved more suitable at the level of the individual word than at the level of the text" (Snell-Hornby, 1995, 80). A key feature of more recent theoretical approaches to translation is their emphasis on the communicative aspects of translation. Translation is seen as a "communicative process which takes place within a social context" (Hatim and Mason, 1990, 3). Instead of seeking for the TL text that is most closely equivalent to the SL input, the goal of translation is to perform an appropriate communicative act in the target community, and the target text is just a means of achieving this goal. Hatim and Mason (1990, 3) point out that doing so requires the study of *procedures* to find out "which techniques produce which effects" in the source and target community.

Interestingly enough, Lefevere and Bassnett (1995, 4) blame the shortcomings of earlier theoretical approaches oriented towards linguistic equivalence on the influence of MT research and its demands for simple concepts that are easy to capture formally. Whether or not this explanation is true, it is striking how firmly even modern SMT techniques are rooted in traditional assumptions of translational equivalence and indeed how apt much of the criticism against such theories of translation is when applied to standard methods in SMT.

Beyond the additional dependencies on pragmatic and cultural knowledge that more recent theories of translation posit, a crucial innovation is that they view translation as an intentional process in its own right. While equivalence-based accounts of translation assume that the best translation of a given input text is somehow predetermined and the translator's responsibility is just to find it, more recent theories recognise that the cultural context and the intended purpose of a translation are not necessarily equal to those of the source text and must therefore be considered as additional variables affecting the desired outcome of the translation process.

3 Word Alignment and Equivalence

The basis of all current SMT methods is the concept of word alignment, which was formalised by Brown et al. (1990; 1993) in the form still used today. Word alignments are objects of elaborate statistical and computational methods, but their linguistic meaning is defined simply by appealing to intuition:

> For simple sentences, it is reasonable to think of the French translation of an English sentence as being generated from the English sentence word by word. Thus, in the sentence pair (*Jean aime Marie*|*John loves Mary*) we feel that *John* produces *Jean*, *loves* produces *aime*, and *Mary* produces *Marie*. We say that a word is *aligned* with the word that it produces.
> (Brown et al., 1990, 80–81)

The authors do not even try to elucidate the status or significance of word alignments in more complex sentences, where the correspondence between source and target words is less intuitive than in the examples cited. In practical applications, word alignments are essentially defined by what is found by the statistical alignment models used, and the issue of interpreting them is usually evaded.

The cross-linguistic relation defined by word alignments is a sort of translational equivalence relation. It maps linguistic elements of the SL to elements of the TL that are presumed to have the same meaning, or convey the same message. The same is true of the phrase pairs of phrase-based SMT (Koehn et al., 2003) and the synchronous context-free grammar rules of hierarchical SMT (Chiang, 2007), which are usually created from simple word alignments with mostly heuristic methods. None of these approaches exploits any procedural knowledge about linguistic techniques and their effects in the source and target community. Instead, it is assumed that each source text has an equivalent target text, possibly dependent on a set of context variables generally subsumed under the concept of *domain*, and that this target text can be constructed compositionally in a bottom-up fashion.

The generation of word alignments is generally governed by two effects: A statistical dictionary or translation table allows the word aligner to spot word correspondences that are very specific in the sense that the occurrence of a particular word in the SL strongly predicts the occurrence of a certain word in the corresponding TL segment. In addition, there is a prior assumption that the word order of the SL and the TL will be at least locally similar, so that the presence of nearby aligned word pairs counts as evidence in favour of aligning two words, even if the link is only weakly supported by the translation table. While the equivalence relation between content words may be strong, it is often more doubtful whether aligned function words really fill exactly the same role in both languages, making these alignments less reliable.

4 Domain as a Catch-All Category

In SMT, the notion of *domain* is used to encode knowledge about the procedural aspects of translation referred to by Hatim and Mason (1990). Domain can be seen as a variable that all the probability distributions learnt by an SMT system are implicitly conditioned on, and it is assumed that if the domain of the system's training data matches the domain to which it will be applied, then the system will output contextually appropriate translations. If there is a mismatch between the training domain and the test domain, the performance of the system can be improved with domain adaptation techniques.

Although there is a great deal of literature on domain adaptation, few authors care to define exactly what a domain is. Frequently, a corpus of data from a single source, or a collection of corpora from similar sources, is referred to as a domain, so that researchers will refer to the "News" domain (referring to diverse collections of news documents from one or more sources such as news agencies or newspapers) or the "Europarl" domain (referring to the collection of documents from the proceedings of the European parliament published in the Europarl corpus) (Koehn, 2005) without investigating the homogeneity of these data sources in detail.

Koehn (2010, 53) briefly discusses the domain concept. He seems to use the word as a synonym of "text type", characterised by (at least) the dimensions of "modality" (spoken or written language) and "topic". Bungum and Gambäck (2011) present an interesting study of how the term is used in SMT research and how it relates to similar concepts in cognitive linguistics. In general, however, the term is used in a rather vague way and can encompass a variety of corpus-level features connected with genre conventions or the circumstances of text use. There is a clear tendency in current SMT to treat all aspects of a text either as very local, n-gram-style features that can easily be handled with the standard decoding algorithm or as corpus-level "domain" features that can conveniently be taken care of at training time.

5 Implications

The use of word-level alignments in SMT is very close to requiring a word-by-word correspondence of the type criticised already by the earliest translation theorists. SMT is a bit more flexible because the dictionaries it uses are created by a relatively unprejudiced statistical algorithm that may include word correspondences a traditional lexicographer would not necessarily agree with even though there is statistical evidence of a correspondence in the training corpus.

The definition of domain as a catch-all corpus-level category is very useful from a technical point of view since it effectively removes all pragmatic aspects from the training procedure itself and replaces them with a single, albeit very strong, assumption of corpus homogeneity. Its downside is that it is quite inflexible. The system cannot adapt easily to different language use in one and the same corpus, for instance when quoted passages differ in style from the surrounding context. Also, it can learn tendencies, but not actual dependencies. As an example, if a target language distinguishes between different levels of formality in its forms of address, domain easily captures which forms are generally preferred in a particular corpus, but it offers no help to decide which form should be selected in each individual case.

In addition, there are circumstances in which the intentionality of the translation process cannot be ignored completely. This happens mostly when the intention of the translation differs from that of the original text. A few such examples are mentioned in the literature. Stymne et al. (2013) describe an SMT system that combines translation with text simplification to cater to target groups with reading difficulties of various types. One of their main problems is the lack of training data having the desired properties on the TL side. However, even if such training data is available, SMT training is not necessarily successful. A case in point is the translation of film subtitles, where the target side is often shortened as well as translated (Pedersen, 2007; Fishel et al., 2012). Anecdotal evidence suggests that MT systems easily learn the length ratio, but truncate the texts in an erratic way that has a negative effect on translation quality.

6 Some Suggestions

Most current approaches to SMT are founded on word alignments in the spirit of Brown et al. (1990). These word alignments have no clear theoretical status, but they can be seen as an embodiment of a fairly traditional concept of translational equivalence. Equivalence in SMT is strongly surface-oriented, and SMT technology has traditionally eschewed all abstract representations of meaning, mapping tokens of the input directly into tokens of the output. This has worked well, demonstrating that much linguistic information is indeed accessible with surface-level processing. However, the SMT system often does not know exactly what it is doing. For instance, based on observational evidence from the training corpus, an SMT system might translate an active sentence in the input with a passive sentence in the output, or a personal construction in the SL with an impersonal construction in the TL without being aware of it. It is difficult to envisage consistently correct translation of complex linguistic phenomena based on such an impoverished representation.

If our goal is to promote progress towards high-quality MT, we should investigate the creation of more expressive cross-lingual representations. The challenge is, then, to do so without compromising the undeniable strength of surface-based SMT. One of its strongest points is its robust descriptive nature that learns as much as possible from data while imposing only very few and general *a priori* constraints. Rather than advocating transfer systems based on specific linguistic theories, we believe that this philosophy should be upheld as much as possible as we explore more expressive transfer representations.

The concept of word alignment works well for content words, and we see no necessity to give it up completely. However, translating function words by mapping them into the TL through word alignments is a more doubtful enterprise, and we suggest that the production of function words should be approached as a problem of generation, or prediction, rather than as a word-level mapping task.

We further believe that it is useful to focus on the correctness of individual structures rather than trying to improve the "average" correctness of an entire text and hoping that individual structures will somehow fall into place automatically. This applies to both translation and evaluation. At translation time, domain adaptation techniques increase the likelihood of correct translations on average, but they do not provide the MT system with any information to support decision-making in particular cases. Therefore, domain adaptation does not appear to be promising as a method to impress a deeper linguistic understanding on SMT; instead, we should strive to overcome the strict dichotomy between word-level and corpus-level modelling and create an additional layer of modelling between the two extremes.

Our stance on evaluation is similar. Aggregating evaluation methods like BLEU (Papineni et al., 2002) give a useful overview of the quality of a translation, but they do not afford specific information and leave too many details to chance. One possible alternative is the creation of test suites with carefully selected examples permitting quick, targeted manual evaluation of specific phenomena in the development phase.

7 Conclusions

Current SMT rests on assumptions of straightforward translational equivalence that oversimplify the complexity of the translation process. Most fundamentally, the central concept of word alignment works well for content words, but is problematic for function words. This leads to problems with controlling the semantics and pragmatics of the translation. Moreover, the intentionality of the translation process is entirely neglected, which causes difficulties particularly when the translation task is combined with some other objective such as text simplification or condensation. This should be borne in mind when designing such translation tasks, but for most applications of SMT, the first problem is clearly more pressing.

The development of new methods in SMT is usually driven by considerations of technical feasibility rather than linguistic theory. This has produced good results, and we expect that it will remain the predominant methodology in the foreseeable future. We consider that it is effective and appropriate to proceed in this way, but from time to time it makes sense to pause and examine the theoretical implications and limitations of the work accomplished, as we have attempted to do for the current standard methods in SMT in this paper.

Acknowledgements

This work was supported by the Swedish Research Council under project 2012-916 *Discourse-Oriented Statistical Machine Translation*.

References

Susan Bassnett. 2011. The translator as cross-cultural mediator. In Kirsten Malmkjær and Kevin Windle, editors, *The Oxford Handbook of Translation Studies*, pages 94–107. Oxford University Press, Oxford.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Computational linguistics*, 19(2):263–311.

Lars Bungum and Björn Gambäck. 2011. A survey of domain adaptation in machine translation: Towards a refinement of domain space. In *Proceedings of the India-Norway Workshop on Web Concepts and Technologies*, Trondheim (Norway).

David Chiang. 2007. Hierarchical phrase-based translation. *Computational linguistics*, 33(2):201–228.

Mark Fishel, Yota Georgakopoulou, Sergio Penkale, Volha Petukhova, Matej Rojc, Martin Volk, and Andy Way. 2012. From subtitles to parallel corpora. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 3–6, Trento (Italy).

Basil Hatim and Ian Mason. 1990. *Discourse and the Translator*. Language in Social Life Series. Longman, London.

Jerome. 1979. Letter LVII: To Pammachius on the best method of translating. In *St. Jerome: Letters and Select Works*, volume VI of *A Select Library of Nicene and Post-Nicene Fathers of the Christian Church, Second Series*, pages 112–119. Eerdmans, Grand Rapids.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton (Canada).

Philipp Koehn. 2005. Europarl: A corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket (Thailand). AAMT.

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

Werner Koller. 1972. *Grundprobleme der Übersetzungstheorie, unter besonderer Berücksichtigung schwedisch-deutscher Übersetzungsfälle*, volume 9 of *Acta Universitatis Stockholmiensis. Stockholmer germanistische Forschungen*. Francke, Bern.

André Lefevere and Susan Bassnett. 1995. Introduction: Proust's grandmother and the thousand and one nights: The 'cultural turn' in translation studies. In Susan Bassnett and André Lefevere, editors, *Translation, History and Culture*, pages 1–14. Cassell, London.

Eugene A. Nida and Charles R. Taber. 1969. *The theory and practice of translation*, volume 8 of *Helps for translators*. Brill, Leiden.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia (Pennsylvania, USA). ACL.

Jan Pedersen. 2007. *Scandinavian subtitles. A comparative study of subtitling norms in Sweden and Denmark with a focus on extralinguistic cultural references*. Ph.D. thesis, Stockholm University, Department of English.

Mary Snell-Hornby. 1995. Linguistic transcoding or cultural transfer? A critique of translation theory in Germany. In Susan Bassnett and André Lefevere, editors, *Translation, History and Culture*, pages 79–86. Cassell, London.

Mary Snell-Hornby. 2010. The turns of translation studies. In *Handbook of Translation Studies*, volume 1, pages 366–370. John Benjamins, Amsterdam.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical machine translation with readability constraints. In Stephan Oepen, Kristin Hagen, and Janne Bondi Johannesse, editors, *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 375–386, Oslo (Norway).