Translating and the Computer 39

16-17 November 2017
One Birdcage Walk, London

Proceedings

Learning from Sparse Data - Meeting the Needs Big Data Can't Reach

Jon D Riding
United Bible Societies
Stonehill Green, Westlea,
Swindon SN5 7TJ
jriding@biblesocieties.org

Neil J Boulton
United Bible Societies
Stonehill Green, Westlea,
Swindon SN5 7TJ
nboulton@biblesocieties.org

Abstract

The vast majority of mainstream MT systems have coalesced around two core technologies, Phrase-Based Statistical Machine Translation (PBSMT) and increasingly Neural Machine Translation (NMT). Both of these technologies have in common the need for very large training data sets. Such data is not available for low resource languages and this is where much of Bible translation takes place. This paper describes a new approach to harnessing the power of machines as Machine Assisted Translation (MAT) engines, supporting the translator in their work from the very start of a project at which point it is likely there is little or no bilingual corpus available. This requires systems with the ability to learn from very small amounts of data and gradually aggregate this knowledge until it is sufficient to support more traditional MT processes. A model for how this might be achieved is presented and the results of early experiments discussed.

1 Introduction

Mainstream MT is largely focussed on synthesis. Systems are designed to translate, at least to first draft, before the human translator's skills are invoked, typically in some form of post-editing. Historically MT systems might be categorised as belonging to one of two types: those which are fundamentally rule-based (RBMT) and those which are heuristic machines of one sort or another. This latter group including various forms of SMT, word or phrase based [Koehn et al, 2003], and more recently NMT systems. All share the characteristic of learning to translate from large example data sets. Of the two the SMT/NMT approach is probably most generally favoured as witnessed by the many implementations of systems based upon generic SMT engines such as Moses and THOT [Ortiz-Martinez & Casuberta, 2014], and the various NMT platforms developed by Google [Wu, 2016] et al.. RBMT continues to contribute not least in the context of hybrid approaches which seek to use the strengths of both RBMT and SMT/NMT approaches [Eisele et al, 2008 & Sanchez-Cartagena et al, 2016] but also in scenarios which are closely controlled and the supporting knowledge bases can be closely tailored to that context. State of the art SMT has coalesced around phrase-based systems.

Both PBSMT/NMT systems have in common a voracious appetite for example data [Shterimov et al, 2017:4] and NMT in particular needs high quality training data to give best results [Nagle, 2017]. Training data sets are commonly measured in millions of documents and whilst NMT is perhaps slightly less hungry than PBSMT in this respect the reality is that a vast data set is needed to train the system. This is analogous to the vast number of exemplars absorbed by a human child as it begins to learn its mother tongue. The principal difference is is that rather than a broad set of exemplars being presented at a single moment in time as is typical for initial training data for PBSMT/NMT a similarly vast set of exemplars is absorbed diachronically over a period of some years and within the wider context of learning that we

Proceedings of the 39th Conference Translating and the Computer, pages 89–97,
London, UK, November 16-17, 2017.

recognise as cognitive development in children [Tomasello, 2008].
This need for enormous sets of training data is of little consequence in the context of mainstream commercial languages where bi-lingual datasets already exist or can be derived from the web. Sub-setting training data for genre improves performance further within that context and the outcome is an excellent set of tools. For so called minority languages where such datasets do not exist and for texts containing many disparate genres and styles the approach is less strong.

2 Translating the Bible

Bible translation is a peculiar problem space. The source text is written in more than one (ancient) language over a period of perhaps 1,600 years with the most recent portions almost 2,000 years old. Not only are we at a considerable distance diachronically from the authors of the text, the target language for a translation of the Bible may be culturally and linguistically distant from the original. The text includes many different genres ranging from narrative to complex constructs designed to emphasise particular concepts or aspects within the text.[1] [App A]. It is very unlikely that much if anything in the nature of training data exists (the translators may well have to being by defining an alphabet). This is not a great scenario for mainstream MT systems and overlaying all these issues is the theological landscape the translation must inhabit both in terms of the particular people, place and time for whom it is prepared and the global context of church and faith.[2] The crucible within which meaning is forged sits at the nexus where the narratives of the text engage with the narrative of the translators and the people they represent. Meaning is instantiated in encounter and it is hard to see how that encounter can be modelled by MT. All of these issues make Bible translators wary of MT as a solution to their task.

3 MT in Bible translation

Many outcomes from MT research during the last twenty years or so in the form of Machine Assisted Translation (MAT) systems have bee embraced by Bible translators and these systems have served Bible translation well. The MAT systems developed for Bible translators focus on analysis rather than synthesis. This objective analysis is then used the inform the work of the translator. Translators have for many years enjoyed the benefit of word-based SMT to analyse the use of key terms in the text, automatic morphological analysis has contributed to spelling checkers and complex pattern recognition systems monitor renderings of items such as proper-names. Crucially, these systems are all entirely language independent, able to operate with any of the 7,000 or so extant world languages without the need for lexica or tables but looking to discern patterns of form, use and meaning within and between texts.
But most of these systems suffer the same limitations as our state of the art PBSMT/NMT systems. They require a lot of training data. The outcome is that they are unable to contribute until a substantial part of the text has been translated, in the case of a New Testament translation perhaps the bulk of the text.

4 Reimagining MAT for Bible translators

The limitations of current machine learning lead to particular problems for Bible translators. The lack of MAT support early in a project leads to many inconsistencies in the text, these in turn contribute to poorer results from MAT systems when they do come online later in a project. To address these limitations we have begun to imagine a new approach to working

1 An example of the complexities which can arise and which are often overlooked by those accustomed to encountering the Bible only in translation can be found at Appendix A.
2 For a thorough exploration of these issues see [Wendland, 2008].

with MAT systems in the context of Bible translation.
Our objective is to bring forward the moment when MAT systems can contribute to a translation, if possible to the very start of a project. We imagine a framework for MAT processing which falls into three main sections:

4.1 Discovery – parseBots

The process of analysing a text must begin with the discovery of entities within the text stream. These may be words, phonemes, morphemes, short phrases or other items but fundamentally they are all patterns which are present in the text. Natural language is inherently structured. The business of learning or reading is, therefore, at its heart an exercise in pattern recognition [Hawkins, 2005 & Kurzweil, 2014].
There are an host of contexts in a text which form patterns. Amongst the key contexts for MAT systems are word formation, clause formation and semantics. Recognising the patterns which arise from these contexts is critical to being able to analyse the text.
We have begun to model a set of processes each of which is targeted towards recognising patterns as they form within a developing text. We call these processes 'parseBots'. These bots can be imagined as peering over the translator's shoulder as he works on a passage of text identifying items in the text stream as the work progresses. Some consider aspects of word-formation such as phoneme stream or morphology, others are concerned with semantics and note close cognates, glosses etc… and others are seeking pattern and order at clause level.
The bots operate on a small pericope of text which we call a 'parse window' [Fig. 1]. This may be as little as a handful of sentences and is unlikely to be as many as 100 sentences. Early experiments indicate that useful analyses can be made from as few as two or three sentences. As identifications are made the patterns found are tentatively tagged for meaning and/or function and added to a collection of parses awaiting confirmation. The parse window follows the translator as he continues to work on the text adding new analyses to the list as new forms enter the window and come under scrutiny. It should be noted that a form that fails to parse on first sight may subsequently be identified as subsequent text enters the parse window and enables an identification to be made. All parses are stored, whether confirmed or not, in the expectation that confirmation may be found in further analyses elsewhere in the text stream. Individual analyses are thus aggregated into a more coherent representation of entities within the developing text.
Setting the extent for the parse window is one of the key research goals for this work. The optimum size is likely to be language and possibly genre dependent and needs to adjust dynamically as it moves through the text, setting the extent such that it exposes the maximum number of parses to the Bots. There is often an assumption that larger extents of text expose more analyses. This leads us to imagine a discovery rate for entities in a text which roughly follows the curve at Fig. 2:

Fig. 1

By this model there is clearly more to be learnt with larger datasets. This is certainly the case but experiment demonstrates that much smaller extents of text can also offer a proportionally rich harvest. A short passage of text with a particular narrative or conceptual focus will expose some entities disproportionately strongly in comparison to their global distribution in the text or language as whole. This characteristic is exploited in a similar fashion to Latent Semantic Analysis [Schone & Jurafsky, 2000]. The consequence of this is that parses which might be overwhelmed in larger extents of text become clearly visible in shorter pericopes.

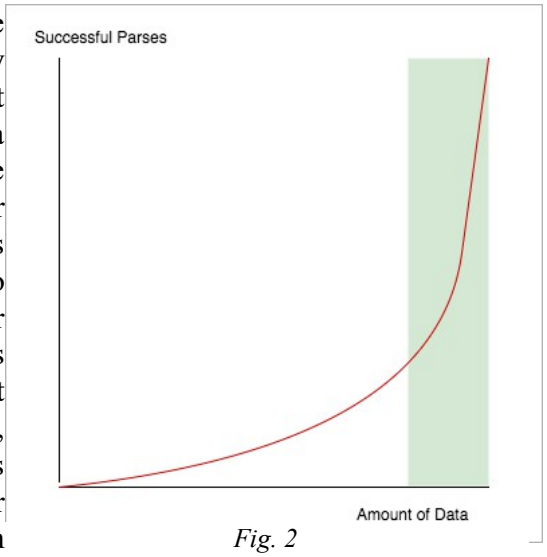

Fig. 2

If this is so we might redraw our discovery graph more like figure 3: to reflect the reality that extent, genre and focus may all contribute to exposing patterns, and so parses, within the text. The precise shape of the curve is likely to be language and context dependent. It is also possible that successful parses from short extent analyses may lessen the depth of the dip as the amount of data increases and bring to the left the moment when analysis of of a larger dataset begins to pay dividends.

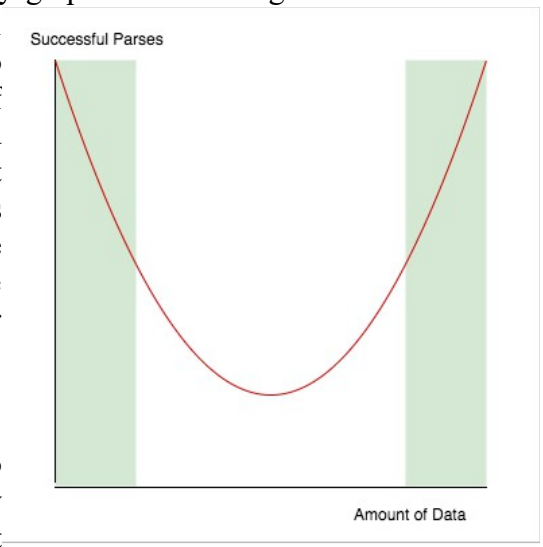

Fig. 3

4.2 Validation – Building a Language Model

Developing a flexible model in which to record this knowledge as it accrues is a key objective for this research. Early experiment suggests that a model based upon surface forms encountered in the text will provide the best framework for recording analyses in preference to attempting to populate predetermined categories of items. Each lexeme encountered by the parse window is stored as part of a developing Language Model (LM). Where parses have been attempted these are stored together with the form. As more forms and parses are added common patterns emerge. A morphological pattern may find support from a number of parses and may in turn generate candidate stem lemmata. If close cognates are identified which confirm the relationship implied by the morphological analysis then the model's confidence in that analysis rises and it may begin to use these parses to drive further analyses as the parse window moves on through the text. It is expected that this aggregation of knowledge within the model will enable local parses to be extended across the model as a whole as patterns emerge which are found to be endemic within the text. It must, however, be recognised that errors, inevitably, creep in as a consequence of the limited processing context. Such errors are a particular concern for a scenario which seeks to exploit limited analyses. If we are to exploit these analyses it is important that we have confidence in them.

4.3 Verification – Confirming the analyses

Given that we cannot rely on having any form of dictionary or grammar for the target language there is only one place we can turn for verification of our parses, the translator. Bible translators are not typically linguists or professional translators as that term is generally understood. They are usually mother-tongue speakers of the target language and, since they represent not just their language community but also the churches within that community, it is likely they will have some measure of theological training. Devising an accessible way to present analyses to translators for assessment is the third important area of research for this project. Initial thinking is that a list of parses awaiting verification will be maintained. As parses acquire a measure of confidence in the LM by aggregation of individual results the proposed identification will be offered to the translator for confirmation or otherwise in the form of a binary question to which the translator can reply only yes or no. For example, a request to confirm that *mundus* and *mundo* refer to the same thing might allow a morphology bot to conclude the possibility of a stem *mund-* with associated morphology *-o, -us*. The subsequent appearance of *mundum* adds *-um* to the morphology and the stem *mund-* can be passed to a glossing bot for confirmation across the wider text.

Translators might choose when they wish to take questions although there may be some merit in maintaining a list of pericope related questions which are presented as the translator finishes a particular passage and whilst the work is fresh in his mind. This represents a departure from the way such confirmation is currently sought. At present, translators cover these kind of checks in sessions lasting hours or even days during which much larger portions of text are reviewed. This is both tedious and time consuming. It is hoped that dealing with such questions piece meal as the work progresses will limit the length of large scale checking sessions and encourage translators to reflect continually on their work as they confirm (or otherwise) the analyses generated by the Bots and the LM.

Over time the LM which is the outcome of this process grows into a database which describes the language encountered in the text and from which resources such as morphology and syntax tables and a bi-lingual dictionary between the source and target text can be compiled. This is exactly the data needed to bring our existing systems such as key terms analysis, morphologically based spelling checks and inter-linear back translation on line at a much earlier stage of the translation.

5 Towards a viable prototype

Our existing systems can provide many of the processes which will power the various bots. If we were to imagine a typical parseBot set as including capabilities in morphological analysis (concatenative and non-concatenative), close cognate recognition, single term glossing, proper-name recognition and some element of part of speech tagging many of these capabilities already exist within the MAT function library that powers the UBS ParaTExt glossing technologies.[3] Re-engineering these systems in the context of sparse data analysis such that parseBots can take advantage of their processing is key to the success of the project.
Constructing a viable Language Model will form a major part of the research needed to realise this proposal. Language Models are more often encountered in RBMT contexts and are typically driven by the expectations of formal linguistics. Language is, sadly, a messy

3 For details of these systems see previous work by the MAT team, much of which has been presented to previous ASLIB/ASLING TC conferences: [Riding (2007), Riding (2008), Rees and Riding (2009), Riding and van Steenbergen (2011), Riding (2012), Riding and Boulton (2016)].

business and experience teaches us that attempting to fit linguistic data drawn from a wide set of languages into a single model based upon abstract classifications is not easy. We propose instead to base our LM on the surface forms encountered in the text together with the parses generated by our processing and the relations implied by those parses. Much of this may well prove very similar to traditional linguistic categories but our objective will be to model the linguistic reality we encounter in the text, rather than to fit the data into predetermined linguistic classes. In addition to establishing a workable data model, the LM will also provide the data for 'global' analyses which attempt to confirm local parse results from the wider data set.

The third area of work facing the team is the need to develop an interaction module to forward confirmation requests to the user and manage their responses. Whilst interactive MT systems are becoming more common these are more commonly used to suggest how a phrase might be completed [Alabau, 2014] rather than to glean information about the text or language. Whether such interactions are best handled 'little and often' or less frequently but in a more structured manner will be another key focus of research as the system is developed.

Acknowledgements

We are indebted to United Bible Societies for their support for this work. Thanks are also due to Oxford Brookes University for continued access to their computing and library services.

References

Alabau V, Buck C, Carl M, Casacuberta F, Garcia-Martinez M, Germann U, Gonzalez-Rubio J, Hill R, Leiva L, Mesa-Lao B, Ortiz D, Saint-Amand H, Sanchis G and Tsoukala C (2014), *"CASMACAT: A Computer-assisted Translation Workbench"*, In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics., 4, 2014. , pp. 25-28.

Eisele A, Federmann C, Saint-Armand H, Jellinghaus M, Herrmann T and Chen Y (2008), *"Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System"*, In Proceedings of the Third Workshop on Statistical Machine Translation., 6, 2008. , pp. 179-182.

Hawkins J and Blakeslee S (2005), *"On Intelligence"* New York, Owl Books.

Koehn P, Och F J and Marcu D (2003), *"Statistical Phrase-Based Translation"*, In Proceedings of HLT-NAACL 2003, Main Papers., 5, 2003. , pp. 48-54.

Kurzweil R (2012), *"How to create a mind",* Viking Penguin.

Nagle P (2017), *"Get the Best from Neural MT with Quality Data"*. KantanMT, URL: https://kantanmtblog.com/2017/08/04/get-the-best-from-neural-mt-with-quality-data/ Retrieved 14:32 1-10-2017.

Ortiz-Martinez D and Casacuberta F (2014), *"The New THOT Toolkit for Fully-Automatic and Interactive Statistical Machine Translation"*, In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics., 4, 2014. , pp. 45-48.

Rees N and Riding J (2009), *"Automatic Concordance Creation for Texts in Any Language"*, In Proceedings of Translation and the Computer 31. IMI/ASLIB.

Riding J (2007), *"A relational method for the automatic analysis of highly-inflectional agglutinative morphologies"*. Thesis at: Oxford Brookes University (MPhil).

Riding J (2008), *"Statistical Glossing, Language Independent Analysis in Bible Translation"*, In Translating and the Computer 30. ASLIB/IMI.

Riding J and van Steenbergen G (2011), *"Glossing Technology in Paratext 7"*, The Bible Translator. Vol. 62(2), pp. 92-102.

Sanchez-Cartagena VM, Perez-Ortiz JA and Sanchez-Marinez F (2016), *"Integrating Rules and Dictionaries from Shallow-Transfer Machine Translation into Phrase-Based Statistical Machine Translation"*, Journal for Artificial Intelligence Research, 1, 2016. (55)

Schone P and Jurafsky D (2000), *"Knowledge-Free Induction of Morphology Using Latent Semantic Analysis"*,

In Proceedings of CoNLL-2000 and LLL-2000. Lisbon , pp. 67-72.

Shterimov D, Nagle P, Casanellas L, Superbo R and O'Dowd T (2017), *"Empirical evaluation of NMT and PBSMT quality for large-scale translation production"*. Thesis at: KantanMT.

Staley, J. (1986) *"The Structure of John's Prologue: Its Implications for the Gospel's Narrative Structure"*, The Catholic Biblical Quarterly, 241-264

Tomasello M (2003), *"Constructing a Language"* Cambridge Mass., Harvard University Press.

Wendland E (2008), *"Contextual Frames of Reference in Translation"* Manchester & Kinderhook, St Jerome Publishing.

Wu Y, Schuster M, Chen Z, Le QV and Norouzi M (2016), *"Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation"*. Thesis at: Google., 10, 2016.

Appendix A – Beyond MT

Here is the opening two verses of St John's Gospel in Greek but transliterated into English characters:

en archē ēn ho logos, kai ho logos ēn pros ton theon,
kai theos ēn ho logos, houtos ēn en archē pros ton theon.

And here is a glossary of the words in those verses:

en	in	*ho logos*	the word	*ton theon*	God
archē	the beginning	*kai*	and	*ho theos*	God
ēn	was	*pros*	with	*houtos*	this (word)

There are three poetic structures used in this short piece of text [Dobson, 2005:6-7 & Staley, 1986] (which in translation is typically presented as a single prose paragraph):

1. Terrace parallelism uses the repetition of key concepts at the end and beginning of adjacent phrases. We can see this happening more clearly in this text if we we highlight them like this:

   en archē ēn ho ***logos***,
   kai ho ***logos*** *ēn pros ton* ***theon***,
   kai ***theos*** *ēn ho* ***logos***,
   houtos *ēn en archē pros ton* ***theon***.

   It is rarely possible to represent this in translation.
2. Threefold repetition adds emphasis or importance to a concept in scripture. Note ***logos*** and ***theos*** above.
3. The writer has not only used terraced parallelism and threefold repetition to highlight 'Word' and 'God', he has also woven a chiastic pattern around the parallelismus:

   A. *en archē* [in the beginning]
   B. *ēn* [was]
   C. *ho logos,* [the word]
   D. *kai ho logos* [and the word]
   E. *ēn* [was]
   F. *pros ton theon,* [with God]
   kai theos [and God]
   E' *ēn* [was]
   D' *ho logos,* [the word]
   C' *houtos* [this (word)]
   B' *ēn* [was]
   A' *en archē pros ton theon.* [in the beginning with God]

   The central point of the chiasm 'F' is the point of emphasis and the closing elements echo the opening elements in reverse order.

Such complexities account for much of the reluctance of Bible translators to embrace typical MT systems.

Appendix B – Initial Experiment

To demonstrate the possibilities of working with very small amounts of data an initial experiment was prepared which used a set of 5 parseBots to analyse the same pericope of John's Gospel in Latin from which the example at Appendix A was drawn. The bots had access to the base text (Greek) which was lemmatised. Beyond this, no information was given other than the text. The text analysed was:

[1]In principio erat Verbum et Verbum erat apud Deum et Deus erat Verbum [2]hoc erat in principio apud Deum [3]omnia per ipsum facta sunt et sine ipso factum est nihil quod factum est [4]in ipso vita erat et vita erat lux hominum [5]et lux in tenebris lucet et tenebrae eam non conprehenderunt [6]fuit homo missus a Deo cui nomen erat Iohannes [7]hic venit in testimonium ut testimonium perhiberet de lumine ut omnes crederent per illum [8]non erat ille lux sed ut testimonium perhiberet de lumine [9]erat lux vera quae inluminat omnem hominem venientem in mundum [10]in mundo erat et mundus per ipsum factus est et mundus eum non cognovit.

The bot set included:

- Close cognate finder
- Morphology analyser
- Lemmatiser
- Proper-name finder
- Glossing engine

The analysis began with a single verse and was then repeated, adding a verse at each iteration and the following hypotheses were queued for verification after each iteration:

1. Cognate: deus, deum?
2.
3. Cognate: ipso, ipsum?
   Cognate: facta, factum?
   Morph: _um?
4. Stem: de_?
   Stem: ips_?
   Stem: fact_?
5. Cognate: tenebrae, tenebris?
6. Name: Iohannes?
   Cognate: non, nomen?
   Gloss: de* = θε*
7. Cognate: omnes, omnia?
   Stem: e_t?
8. Cognate: hic, hoc?
   Cognate: ille, illum?
9. Cognate: hominem, hominum?
   Stem: homin_
10. Cognate: facta, factum, factus?
    Gloss: fact* = ποι*
    Cognate: mundo, mundum, mundus?
    Stem: mund_?
    Gloss: mund* = κοσμ*

Results marked in green are analyses confirmed by more than one bot process. These are forwarded to the user for verification via the interaction module.

Of the remainder, all but the non/nomen cognate are valid and we can expect that to be dismissed by subsequent processing.

Particularly pleasing is the *hic/hoc* cognate which illustrates the power of non-concatentive morphology analysis

The *e_t* stem is also of interest. At first sight this is nonsensical but the data that support it are in fact *est/erat*; cognate forms of the Latin verb to be.

This is a rich harvest from so small a data set.