

SALT cymru_

Document 6

An overview of the Moses machine translation system, and its possible enhancement for use in Wales in a pre-competitive research stage

**Prepared by the
Language Technologies Unit (Canolfan Bedwyr), Bangor University**

April 2008



Llywodraeth Cynulliad Cymru
Welsh Assembly Government

This document was prepared as part of the SALT Cymru project, funded by the Welsh Assembly Government under the Knowledge Exploitation Fund's Knowledge Exchange Programme, reference HE 06 KEP 1002

Overview

By reviewing contemporary literature on Machine Translation (MT), current best practice was identified, and a system conforming to these practices was evaluated, with the requirements of SMEs in Wales in mind. We found that, given a suitable bilingual text corpus, very effective machine translation can be achieved. Further research which could be of immediate practical benefit was identified, and relevant issues noted accordingly.

Machine Translation Paradigms

There are three major classes of machine translation.

Statistical Machine Translation (SMT): a bilingual text corpus is analysed to produce a statistical model of the mapping from a source language to a target language. Subsequently, given text in the source language, a most likely equivalent in the target language is found according to this model. SMT is currently the paradigm in which the majority of MT research and development is being undertaken.

Rule-based Machine Translation: a system of lexical, grammatical and reordering rules is created for a source-language/target-language pair. The rules are then applied to subsequent source text to produce translated output.

Example-based Machine Translation (EBMT): a bilingual text corpus is used directly for comparison against source text, and case-based reasoning is applied to create a translation.

Purported EBMT systems are often in fact hybrids of the above approaches to some extent, and systems exist using various combinations of these techniques and others.

Evaluation of an SMT system

One of the current leading SMT systems is Moses, a factored phrase-based beam-search decoder, a free, open-source project licensed under the LGPL¹. An online demonstration of Moses is available², using one of the most popular and comprehensive freely available corpora, the European Parliament Proceedings Parallel Corpus (Europarl)³.

Moses was evaluated using a bilingual English-Welsh text corpus. Other language pairs would also be useful to SMEs in Wales, in dealing with minority language communities as well as international communication. But since SMT uses a bilingual text corpus, and not specific rules about the languages in question, most of the issues would be similar in all these cases.

¹ <http://www.statmt.org/>

² <http://demo.statmt.org/webtrans/>

³ <http://www.statmt.org/europarl/>

Unfortunately the Europarl corpus does not include the Welsh language because Welsh is not an official working language of the EU. The availability of an appropriate corpus is of primary importance for SMT, and in our opinion, likely to be the limiting factor in the development of effective MT systems, particularly for lesser-resourced languages.

Therefore, for evaluation purposes, a corpus was assembled from the Welsh Statutory Instruments, which are legislation published in English and Welsh in accordance with the Government of Wales Act 1998⁴; this makes a useful test system because legislative language is highly regular and specialised. The more comprehensive Welsh Assembly Proceedings corpus assembled by David Talbot contains actual spoken language and as such would likely produce better results in a variety of more general contexts. The Cronfa Electronig o Gymraeg⁵ assembled by the Language Technologies Unit at Bangor University is not appropriate for this purpose because it is monolingual (but see below).

Moses treats each sentence independently, so sentence boundaries were identified and the English and Welsh versions matched up, using a custom script written in the Python programming language. Note that this sentence-by-sentence approach implies that a corpus whose translations are relatively non-literal (with sentences being combined or split) will be drastically less effective.

The statistical models were generated from the corpus by a computer program running overnight. Note that more time would be required for a larger corpus, and that this could conceivably have implications for iterative development, where the software developer repeatedly makes small changes to the system and then tests it. If the developer must wait several hours to try out each small modification, the overall project could take a long time; therefore, one of the software development techniques which avoid this problem would have to be used.

Analysis of Translation Quality

The quality of translation obtained is highly dependent on the bilingual corpus used -- in this case, legislation, which is written in specialised formal language. The system performs well when translating text of this type. Here is a hypothetical sentence from a non-existent Regulation, using many common phrases:

The Welsh Ministers make the following Regulations in exercise of the powers conferred on the Secretary of State by section 1 of the Local Government Act 1972[1] and now exercisable by the National Assembly for Wales.

And here is the suggested translation:

Mae Cynulliad Cenedlaethol Cymru yn gwneud y Rheoliadau canlynol drwy arfer y pwerau a roddwyd i'r Ysgrifennydd Gwladol gan adran 1 o Ddeddf Llywodraeth Leol 1972[1] ac sy'n arferadwy bellach gan Gynulliad Cenedlaethol Cymru.

In this case the translation is perfect. In other cases, there are errors but the suggestion is a useful starting point and could serve as an effective enhancement to a human translator, significantly increasing the rate at which translation work can be performed.

⁴ <http://www.opsi.gov.uk/legislation/wales/w-stat.htm>

⁵ http://www.bangor.ac.uk/~cbs204/ceg/newidiadau_i_dagiau_ceg.html

At the opposite extreme, the system has almost no idea how to translate the following sentence:

Sheep only eat grass

The output is:

Dafad eat grass yn unig

The words “eat” and “grass” actually occur repeatedly *in English* in the “Welsh-language” legislation, because it quotes monolingual legislation from the UK Parliament which is being amended⁶. Text of this type should be omitted from the corpus to avoid this happening.

We conclude that, given a suitable corpus, SMT can generate good-quality English-Welsh translations and could be very useful to a human translator.

Areas for Further Research

It is clear that efficient SMT requires a large corpus of high-quality bilingual text. The system is most effective when used on material similar to that in the corpus. It may be useful to have several corpora for different subject areas. There is a great deal of bilingual text available from public institutions in Wales, which could be used given appropriate resources to collate and prepare the material in the appropriate format. Both manual collation and technological solutions such as automatic alignment (matching up sentences) would be of benefit. **Assembling an appropriate bilingual corpus is the major factor in achieving effective English-Welsh machine translation.**

English-Welsh MT tools would be more useful if integrated into translation memory frameworks, a number of which are in widespread use in Wales, both proprietary and free/open-source⁷. When a document is being translated, the machine translation can then conveniently act as a starting point for further refinement by a human translator. This is a step beyond current practice in Wales, whereby translated texts are stored and suggested if they occur again; MT can make new suggestions for phrases which it has never seen before.

Translation involving other languages would also be useful to SMEs in Wales, e.g. Polish, Chinese. In many instances the typical usage might be somewhat different; for instance an imperfect machine translation might actually be of direct use to a person who understands little English, though caution should be applied not to create unrealistic expectations⁸.

⁶ See, for instance, <http://www.opsi.gov.uk/legislation/wales/wsi2005/20051156w.htm#7>

⁷ e.g. Trados (<http://www.trados.com>), Wordfast (<http://www.wordfast.net>), OmegaT (<http://www.OmegaT.org>)

⁸ A number of very public Welsh mis-translations (e.g. See <http://news.bbc.co.uk/1/hi/wales/5341646.stm>) have resulted from a misleadingly marketed program which performs rudimentary and erroneous word-by-word translation (<http://www.tranexp.com/#InteractiveTrananchor>)

Effective tools⁹ already exist for identifying Welsh parts of speech (and, for instance, recognising that “cael” (get) and “cafodd” (got) are different forms of the same word), and output from these could enhance Moses's word-recognition capabilities. The monolingual CEG corpus could also be used to improve word recognition.

Due to time constraints, many of Moses's default “models” were used. One example is that translations are preferred where the word order of the translated sentence is close to that of the original text. We suspect that there may be better models for Welsh-English translation, because of significant differences in word order (VSO/SVO, pre/post-positional adjectives, periphrastic verb forms etc).

⁹ The lemmatizer component of Cysill, developed by the Language Technologies Unit, Canolfan Bedwyr, Bangor University