Phrase-based Statistical Machine Translation between English and Welsh

Dafydd Jones, Andreas Eisele

Dept. of Computational Linguistics
Saarland University, PO Box 151150
D-66041 Saarbrücken, Germany
dafyddj@gmail.com, eisele@coli.uni-saarland.de

Abstract

This paper shows how a baseline phrase-based statistical machine translation (SMT) system can be set up for translation between English and Welsh, a UK language spoken by about 610,000 people, using well-documented and freely available tools and techniques. Our results indicate that the achievable performance for this language pair is among the better of those European languages reported in Koehn (2005). We argue that these preliminary results should be seen as a first step towards hybrid systems where linguistic knowledge from lexical and morphological resources are combined with additional terminology and stochastic preferences acquired from existing translations.

1. Introduction

As demonstrated in Koehn (2005), phrase-based statistical machine translation (SMT) can lead to interesting levels of translation quality for many language pairs if parallel corpora of sufficient size can be used for training. Results are particularly notable for target languages that do not make use of extensive morphology.

Compared with older models of SMT, where decoding (translation) methods were based directly on the mathematical models used for computing the word alignments, the phrase-based variant of SMT has the advantage of being flexible in the size of the building blocks it uses. Re-use of long subsequences of sentences leads to the high precision known from the use of translation memories (TMs). In contrast to TMs, where recall can fall dramatically when applied to texts written from scratch, phrase-based SMT shows a much softer degradation in these cases.

The availability of relevant tools (Koehn, 2006) made possible the organisation of a project seminar during the Summer semester, 2005 at Saarland University to investigate the use of these tools in the development of at least baseline SMT functionality for a number of different language pairs. Whereas most of these systems were built using the Europarl corpora (Koehn, 2002), and did not explore further than the 110 language pairs covered in Koehn (2005), our work involving the minority language of Welsh shows that this particular approach is equally applicable when one member of the language pair has little conventional language technology available to those wishing to conduct state-of-the-art computational linguistic research.

2. Corpus collection

Of crucial importance for any statistical analysis of language is the availability of a sufficiently large, high-quality corpus. Our corpus is collected from the online version of the Record of Proceedings of the Plenary Meetings of the National Assembly for Wales.

The Welsh Assembly is an elected body representing almost 3 million people and is responsible for developing policy and allocating funds made available to it from the UK Treasury. It consists of 60 members, who held their first meeting in May, 1999. Plenary meetings are held twice a week, and consist of questions and debate. Meetings are bilingual, and a speaker may speak in the language of their choice.

A verbatim report is made available to the public within 24 hours of any given meeting, with a 5-day, fully translated, official version made available on the Assembly website <http://www.wales.gov.uk/organipo/>. Though its publications are subject to Crown copyright protection, the Assembly has a policy of open access to information, and its records are free to use and can be reproduced under waiver conditions. The waiver conditions allow reproduction and distribution for any purpose and in any medium as long as the Record remains accurate and Crown copyright is acknowledged.

Collecting the source material for the corpus was simply a matter of crawling the relevant sections of the Assembly's website. The proceedings are presented as HTML in a 2-column table format, where each row of the table contains a turn of a speaker, one column for each language.

Of slight inconvenience is the fact that the left column always represents the speaker's utterance in its original language. A number of Assembly members choose to make their comments in Welsh, so the contents of the left column arbitrarily switches from English to Welsh. An ad-hoc language guesser was constructed, using data from the CPAN Perl module Language::Guess (Ceglowski, 2004), to identify Welsh and English paragraphs.

After sentence segmentation, corresponding Welsh/English paragraphs that differed in number of sentences were rejected. Inspection revealed that after this step corresponding sentences within speaker turns were aligned between languages. A final cleaning step was performed to remove a few garbage lines, convert HTML entities to ASCII equivalents and tokenize words and punctuation (apostrophes and hyphens within words were preserved). Table 1 shows some measurements of the final bilingual, sentence-aligned corpus.

	Welsh	English
Sentences	510,813	
Tokens	10,760,861	10,703,378
Types	65,219	49,719

Table 1: Corpus measurements after processing

Given the high-quality and size of this corpus, we are happy to make it available to other interested researchers.

Even in a monolingual form, it currently represents the largest freely-available Welsh corpus. Contact <dafyddj@gmail.com> for more details.

3. Software

Our work was made possible by the availability of software tools that cover the whole system development cycle from initial word-alignment and phrase table training to decoding and evaluation.

3.1. GIZA++

GIZA++, developed by Franz Josef Och (2003), is used to calculate word alignments between corresponding bilingual sentences according to refined statistical models. A detailed description of the design of the software can be found in Och and Ney (2003).

GIZA++ is itself an extension of the original GIZA software developed as part of the 1999 Summer workshop hosted by the Center for Language and Speech Processing at Johns Hopkins University (CLSP, 1999). The software is released under the GNU Public License.

3.2. Pharaoh

Phillip Koehn's Pharaoh (2004) system consists of scripts for extracting phrases from word-aligned sentences (provided by GIZA++, for example), and a decoder, which actually performs translation of an input sentence, given an appropriate translation model and a target language model. In this case, the translation model is a phrase-table that contains a set of phrase correspondences in the form of a foreign phrase, a native phrase, and a conditional probability that one could be the translation of the other.

Pharaoh is provided (Koehn, 2004b), without source code, for non-commercial purposes by the University of Southern California.

3.3. SRI Language Modeling Toolkit

SRI International provide a toolkit (SRI, 2006) for building and applying n-gram based language models under an open-source community licence that allows not-for-profit use and requires any code changes to be shared with other users. The Pharaoh decoder works with SRI language models.

4. Training and Decoding

Philipp Koehn provides a script, 'train-phrase-model.perl', as part of the Pharaoh package that automates the training process, using the above software. We split our corpus into a training set of 460,813 sentences, holding out 50,000 sentences for potential tuning and testing. Using this training data we generated bi-directional phrase tables of around 20 million phrases each (approx. 1.8 Gb in size on disk).

Due to the problem of loading and storing such a large amount of data in memory, sentence translation is accomplished via a filter script, that extracts a subset of phrases from the full phrase tables that account for actual phrases found in the source sentences. Nevertheless, this filtering incurs a time penalty, and data structures can still take over a minute to load into memory, with subsequent translation taking between 5 and 10 seconds per sentence.

5. Evaluation

Systematically measuring similarity between MT results and a human reference translation has become quite popular in the last five years (Papineni et al., 2002), but the metrics used for these comparisons, such as the BLEU score, are typically very superficial and do not allow qualified statements on absolute translation quality or comparison between systems across widely different architectures (Callison-Burch et al., 2006).

On the other hand, automatic evaluation has the advantage of being cheap, both in time and resources, and is therefore appropriate for measuring progress or regression during the development of a given system. In that sense, our measurements are based on BLEU, due to lack of time for more extensive manual investigations, and should only be viewed as a very first step towards a meaningful evaluation.

We picked a 5000 sentence test set, from previously held out data, as a basis for our measurements. Table 2 shows the BLEU scores measured on translations from Welsh to English and vice-versa. Figure 1 shows an example set of translations.

English to Welsh
Source ' iaith pawb ' clearly states that the availability of education through the medium of welsh has increased steadily in recent years , and that that is an aim your government wants to encourage
Translation mae ' iaith pawb ' yn datgan yn glir bod y ddarpariaeth o addysg drwy gyfrwng y gymraeg wedi cynyddu raddol yn ystod y blynyddoedd diwethaf , a bod eich llywodraeth yn dymuno annog nod
Welsh to English
Source mae ' iaith pawb ' yn datgan yn glir fod y gallu i gael addysg drwy gyfrwng y gymraeg wedi cynyddu'n gyson yn y blynyddoedd diwethaf , a bod eich llywodraeth yn dymuno hybu'r amcan hwnnw
Translation ' iaith pawb ' states clearly that the able to receive their education through the medium of welsh has steadily increased in recent years , and that your government wants to promote that objective

Figure 1: Examples of a spoken utterance and its corresponding translations.

From Welsh	Into Welsh
40.22	**36.17**

Table 2: BLEU scores for Welsh-English translation calculated over 5000 test sentences

To give some context to these measurements, the highest and lowest BLEU scores reported in Koehn (2005) are 40.2 for translating Spanish to French, and 10.3 for Dutch to Finnish. It should be noted that these systems were developed and tested on a different corpus from a different domain.

6. Other work

Our work is not the first reported instance of Welsh to English statistical machine translation. Phillips (2001) reports on the implementation of software to construct stochastic translation models from bilingual sources. In an interesting approach, he used the Bible as training text for building his system, extending this with a morphological component and a bilingual dictionary to compensate for the limited vocabulary of the Bible.

Due to a lack of an independent corpus, our Welsh language model was generated from our approx. 10 million word training corpus. Kevin P. Scannell's An Crúbadán project (Scannell, 2004) has collected a Welsh corpus of 95 million words. His software crawls the web, specifically collecting texts in minority languages, bootstrapping further crawls by using seed text as search terms to discover more web pages in the target language. We hope to investigate the use of this internet corpus as a basis for the language model required for translation into Welsh.

7. Conclusions

We see our SMT system as a first step towards machine translation for Welsh. We are convinced that better quality and coverage can be achieved when linguistic knowledge, such as rule-based morphological analysis and parsing is included in the process. Whereas this may look straightforward on the English side, a lack of language technology resources for Welsh is a hindrance.

However, we see the option to use the existence of large amounts of parallel texts and high-quality word alignments to project (parts of) linguistic analyses along these alignments to the other language, thus bootstrapping linguistic knowledge for Welsh in a manner that avoids the expense of treebanking but still promises higher quality than fully unsupervised approaches. We plan to use techniques and results from the Ptolemaios project (Kuhn, 2004) for a further exploration of this perspective.

We note that there is currently some interest in the commercial development of Welsh/English machine translation. A report commissioned by the Welsh Language Board (Somers, 2004) makes the recommendation that initial work in this area should be focused on the development of an SMT system capable of producing low-quality but usable translations. We hope our work shows that this is indeed a realistically achievable goal.

8. References

Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *EACL-2006 (to appear).*

Ceglowski, M. (2004). Language::Guess (version 0.01). CPAN, accessed April, 2006, <http://search.cpan.org/~mceglows/ Language-Guess-0.01/Guess.pm>.

Center for Language and Speech Processing (1999). The EGYPT Statistical Machine Translation Toolkit. Johns-Hopkins University, accessed April 2006, <http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/>.

Koehn, P. (2002). Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Unpublished draft, MIT, accessed April 2006, <http://people.csail.mit.edu/~koehn/publications/ europarl.ps>.

Koehn, P. (2004). Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation in the Americas*, Lecture Notes in Computer Science. AMTA, Springer.

Koehn, P. (2004b). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. ISI, accessed April, 2006, <http://www.isi.edu/ licensed-sw/pharaoh/>.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*.

Koehn, P. (2006). Statistical Machine Translation. Accessed April 2006, <http://www.statmt.org/>.

Kuhn, J. (2004). Experiments in parallel-text based grammar induction. In *42nd Annual Meeting of the Association for Computational Linguistics*. ACL.

Och, F. J. (2003). GIZA++: Training of statistical translation models. Accessed April, 2006, <http://www.fjoch.com/GIZA++.html>.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*. ACL.

Phillips, J. D. (2001). The Bible as a basis for machine translation. In *Proceedings of Pacling 2001*. Pacific Association for Computational Linguistics.

Scannell, K. P. (2004). Corpus building for minority languages. Saint Louis University, accessed April 2006, <http://borel.slu.edu/crubadan/>.

Somers, H. (2004). Machine translation and Welsh: The way forward. Technical report, The Welsh Language Board.

SRI International (2006). SRILM - The SRI Language Modeling Toolkit. Accessed April 2006, <http://www.speech.sri.com/projects/srilm/>.