PBML

logo

The Prague Bulletin of Mathematical Linguistics

NUMBER ??? JANUARY 2009 1–10

apertium-cy - a collaboratively-developed free RBMT system for Welsh to English

Francis Tyers, Kevin Donnelly

Abstract

apertium-cy (`http://www.cymraeg.org.uk`) is a rule-based "gisting" machine translation system for Welsh to English, with both engine and data released under the GPL. We summarise the development of *apertium-cy*, evaluate its output, and discuss the advantages of a collaborative development model combined with rule-based MT for marginalised languages.

1. The Apertium platform

apertium-cy is a "gisting" machine translation system for Welsh to English, based on the Apertium machine translation platform.[1] The platform was originally aimed at the Romance languages of the Iberian peninsula, but is now being adapted for other languages (such as Basque, and languages from the Celtic group – Welsh, Irish, Breton), with much of the work on new languages being pursued by volunteers, following the increasingly common collaborative development model used for free[2] and open-source software. Apertium is licensed under the Free Software Foundation's GNU General Public License,[3] and all the software and data for the 17 supported language pairs (and the other pairs in development) is available for download from the project website.

Apertium follows a shallow-transfer approach, and is very fast. Finite-state transducers (Garrido-Garrido-Alenda and Forcada, 2002, Roche and Schabes, 1997) processing up to 40,000 words per second are used for lexical processing, first-order hidden Markov models (HMM) are used for part-of-speech tagging, and multi-stage finite-state based chunking for structural transfer.

[1] `http://www.apertium.org`

[2] We follow the definition of "free" used by the Free Software Foundation - `http://www.fsf.org`.

[3] `http://www.fsf.org/licensing/licenses/gpl.html`, accessed 12/12/2008.

Please cite this article as: Francis Tyers, Kevin Donnelly, *apertium-cy* - a collaboratively-developed free RBMT system for Welsh to English. The Prague Bulletin of Mathematical Linguistics No. ???, 2009, 1–10.

The software behind the platform is implemented as a standard UNIX pipeline, with each stage in the translation having a separate C++ program. Communication between each stage uses piped text streams. XML-based formats are used to encode the linguistic data, which are then compiled into the high-speed formats used by the engine. Further details are given in (Armentano-Armentano-Oller et al., 2006), and on the project website.

2. Background

The increasing economic and cultural importance of information technology poses a threat to marginalised languages. Unless they have an ICT presence, along with a reasonable range of the tools that computer-users take for granted (e.g. spelling and grammar-checkers, dictionaries and thesauruses, etc.), they run the risk of being shut out of this whole sector of modern life. The result will be a further decline in their status and usage (Crystal, 2000).

In the case of Welsh, which could be said to be the healthiest of the Celtic languages, with nearly 600,000 speakers (almost 21% of the population),[4] requests had been made over a period of years for leading software suppliers to produce Welsh versions of their software, but little progress had been made. Free software offered a way around this problem.

In 2003, the second author launched a voluntary initiative to translate software used on the GNU/Linux operating system into Welsh. Language tools that majority languages take for granted would have been useful to correct mistakes and maintain quality, but unfortunately, where such tools existed, they were not available under a licence which would enable them to be distributed with the software being translated (the exception being a Welsh spellchecker which was released some time later for the Firefox browser).

The only alternative was to create these tools from scratch. This involved delaying the translation project, and working in turn on a dictionary,[5] verb inflector,[6] and a grammar-checker based on *An Gramadóir* (Scannell, 2008).

3. Development of *apertium-cy*

When we started extending Apertium to create a gisting system for Welsh by using this free data, a major issue to be tackled was revising its dictionary format to deal with perhaps the defining feature of the Celtic languages – mutation (alteration of initial phonemes, usually having a morphological significance). An example would be the lemma *tad* (father), which can appear as *dad* (*ei dad* – his father), *nhad* (*fy nhad* – my father) and *thad* (*ei thad* – her father). For further details, see (Ball and Müller, 1992) and (Stewart, 2004).

It also became clear that there were two main issues with the disambiguation stage in the Apertium platform with respect to Welsh. The first was that the HMM-based POS tagger was

[4] Table WLP01 in: `http://www.statistics.gov.uk/downloads/census2001/Report_on_the_Welsh_language.pdf`, accessed 12/12/2008.

[5] `http://www.eurfa.org.uk`

[6] `http://www.konjugator.org.uk`

not able to take advantage of the disambiguation properties of the mutations, and the second was that the accuracy of a tagger trained in an unsupervised manner (without a tagged corpus) was below what was expected.

The following phrase gives an example of mutation having an effect on disambiguation: *Mae'r modd y mae gweithwyr mudol...* (*The means by which migrant workers...are...*). Here, *modd* could be interpreted as either an unmutated form of the noun *modd* (means) or as a nasally-mutated form of the noun *bodd* (pleasure). Linguistically there is no ambiguity – a nasal mutation would never follow the definite article, *[y]r*.

Since there was no tagged corpus of Welsh available under a free licence with which to train the tagger in a supervised manner, we examined other options for improving disambiguation performance. Rather than add this functionality directly to the Apertium tagger, the first author undertook a review of free software which might meet our needs here, and settled on Constraint Grammar[7], being developed by researchers at the University of Southern Denmark – further details are available in (Karlsson et al., 1995), and on the webpage referenced. It proved possible to integrate this software relatively easily, and a number of CG rules were written to disambiguate the Welsh text before it was fed into the statistical tagger.

Version 0.1 of *apertium-cy* contains approximately 10,000 lemmas and 150 grammatical rules, and has up to 94% coverage on large Welsh corpora.[8] Work continues on 0.2, with the key target being an initial version of an English to Welsh translator.

4. Evaluation

Previous work reported for Welsh machine translation includes (Phillips, 2001) and (Jones and Eisele, 2006). (Somers, 2004), in a report commissioned by the Welsh Language Board on a possible strategy for MT in Welsh, suggests (section 4.1) that one of the initial steps should be "a small Welsh-English...system for gist translation of Welsh documents". We would argue that *apertium-cy* constitutes such a system.

4.1. Quantitative

In order to provide a comparison with the other published paper on the subject, (Jones and Eisele, 2006), we evaluated the quality of the translations using three common metrics: word error rate (WER), position-independent word error rate (PER), and the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002).

The first two metrics (WER and PER) provide a measure of post-edition effort – that is, they measure how much of the text needs to be changed to achieve a translation of publishable quality.

The third (BLEU) is a popular evaluation metric within the machine translation community, and we include it to give results comparable with the system described in (Jones and Eisele,

[7] `http://beta.visl.sdu.dk/constraint_grammar.html`, accessed 12/12/2008.

[8] Coverage for the PNAW corpus – see section 4.1 below – is 94%, and for the Welsh Wikipedia, 86%.

2006). It should be noted that BLEU scores are not necessarily appropriate for comparing systems of different types, and have a tendency to penalise rule-based systems (Callison-Callison-Burch, Osborne, and Koehn, 2006, Labaka et al., 2007).

Two corpora were used for evaluation. The first corpus consisting of 318 sentences (5,492 words) was selected at random from the Welsh Wikipedia - we have made this corpus available for download.[9] These sentences were translated and then post-edited by a speaker of both languages. We then calculated the WER, PER and BLEU scores of the system's translations against the post-edited references. It is worth noting that (a) sentences were filtered such that each one is covered fully by the dictionaries, with the result that this test does not give an indication of system vocabulary coverage, and (b) using BLEU against a post-edited reference is a novel methodology and may lead to higher scores.

The second corpus consisted of a 10% sample (50,000 sentences) of the Welsh–English Proceedings of the National Assembly for Wales (PNAW) corpus.[10] For this corpus, the sentence sample was translated, and then the scores were calculated between the translated sentences and the reference translations.[11] As with (Jones and Eisele, 2006), unknown words were left in the translations.

The results of these evaluations are given in Table 1 – bold face represents the highest score.

	WER	**PER**	**BLEU**
Wikipedia true-case	55.78	30.59	30.70
Wikipedia lowercased	**53.40**	**27.22**	**32.21**
PNAW true-case	65.99	35.44	15.12
PNAW lowercased	**64.94**	**34.35**	**15.68**

Table 1. ***WER, PER and BLEU metrics for the two corpora***

As expected, considering the system is intended for "gisting translation", the scores for the post-edition task indicate that the system currently produces sentences with too high an error rate to be usable for professional post-editing.

There is less difference between the PER scores for each corpus than there is for the WER and BLEU scores. This probably indicates (unsurprisingly given the differences in sentence structure between Welsh and English) that more needs to be done on word-reordering, which this particular metric does not take into account.

It is also surprising that the BLEU score for the PNAW corpus is substantially lower than for the Wikipedia one, while the scores for WER and PER are less divergent. All of the BLEU

[9] `http://xixona.dlsi.ua.es/~fran/welsh/cy-test-corpus.tar.gz`

[10] See (Jones and Eisele, 2006), available from `http://xixona.dlsi.ua.es/corpora/`.

[11] It is generally recommended that three references are used as a minimum for calculating BLEU scores. Unfortunately three references were not available, and (Jones and Eisele, 2006) reports using a single reference translation.

scores are lower than the 40.22 for Welsh to English reported in (Jones and Eisele, 2006), but given the variation in the scores we suspect, along with the previously mentioned authors, that BLEU is not a particularly good metric for comparing unrelated MT systems.

4.2. Qualitative

The targets for *apertium-cy* 0.1 included the aim that "sentences of up to 5 words should be translated reasonably well from Welsh to English",[12] since this will allow a large number of short, conceptually-simple sentences to be translated. This has been met quite comfortably, as the following examples (with *apertium-cy* output in line b) show:

(1) a. *Mae'r gath yn yr ardd, ond mae'r ci yn y cae.*
b. The cat is in the garden, but the dog is in the field.

(2) a. *Mi welodd y dyn y bachgen yn dod allan o'r siop.*
b. The man saw the boy coming out of the shop.

(3) a. *Mi fydd y trên yn hwyr yfory, oherwydd bydd y cwmni yn gweithio ar y lein.*
b. The train will be late tomorrow, because the company will be working on the line.

Other sentences do not come across as well-formed English at present, and are therefore not suitable for dissemination, but the meaning is perfectly clear:

(4) a. *Bydd rhaid i ti frwsio dy ddannedd cyn mynd i'r ysgol.*
b. *Necessity will be to you brush your teeth before go to the school.
c. You'll have to brush your teeth before going to school.

(5) a. *Mi gafodd yr adroddiad ei olygu gan y Pwyllgor Seneddol.*
b. *The report got edit with the Parliamentary Committee.
c. The report was edited by the Parliamentary Committee.

apertium-cy 0.1 has been tested on official statements, novels, newspaper articles and non-fiction, and also seems to work on older texts provided the spelling is modernised, as in this example from 1865:

(6) a. *Ond oherwydd na ellid disgwyl ond cylchrediad lleol i lyfr o'r fath, ac oherwydd nad oedd gennym ninnau arian i'w gwario mewn ymgymeriad felly, ofnwyd yr anturiaeth.*
b. *But because nor could expect but local circulation to book of the type, and because was not with us us a money to spend it in undertaking so, the adventure were feared.

[12] `http://wiki.apertium.org/wiki/Welsh_to_English`

c. But since only a local circulation for a book of this kind could be expected, and because we ourselves did not have the money to spend on such an undertaking, the venture raised fears.

The web interface at `http://www.cymraeg.org.uk` gives other examples, and allows users to enter their own text in order to come to their own conclusions about translation quality.

5. Shortcomings

The main shortcomings of *apertium-cy* 0.1 fall into three main areas: phrase delimitation, treatment of subordinate clauses, and lexical selection.

5.1. Phrase delimitation

Because Apertium follows a shallow transfer approach, and does not include a full parser, it can be difficult to delimit phrases in such a way that they can be handled as a block when the target language requires this. Consider the following series of examples, based on the standard Welsh genitival construction, where the sequence *noun1 + def.art + noun2* is equivalent to the English *the noun1 of the noun2* or *the noun2's noun1*:

(7) a. *cath* *ddu*
[$_{\text{NP}}$ noun + qual]
b. black cat

(8) a. *cath* *y* *meddyg*
[$_{\text{NP}}$ noun + def.art + noun]
b. the cat of the doctor
c. the doctor's cat

(9) a. *cath* *ddu* *y* *meddyg*
[$_{\text{NP}}$ noun + qual + def.art + noun]
b. *black cat the doctor
c. the doctor's black cat

(10) a. *cath* *ddu* *fawr* *y* *meddyg*
[$_{\text{NP}}$ noun + qual + qual + def.art + noun]
b. *big black cat the doctor
c. the doctor's big black cat

(11) a. *cath* *merch* *y* *meddyg*
[$_{\text{NP}}$ noun + [$_{\text{NP}}$ noun + def.art + noun]]
b. *daughter cat the doctor
c. the doctor's daughter's cat

(12) a. *cath* *ddu* *merch* *y* *meddyg*
[$_{\text{NP}}$ noun + qual + [$_{\text{NP}}$ noun + def.art + noun]]
b. *daughter black cat the doctor
c. the doctor's daughter's black cat

(7) and (8) are well-formed English, and (9) and (10) are only missing a definite article and the genitival *of*. However, in (11), the sub-phrase *merch y meddyg* should be translated along the lines of (8) as *the daughter of the doctor*, and left in position. Instead, *merch* is treated as a qualifier like *ddu*, and shifted, as can be seen when comparing (12) and (10). Most instances of problematic phrase delimitation are more complex than this, but we are confident that adding more rules and refining existing ones, along with the planned additon of a basic parser to Apertium some time in 2009, will resolve many of the issues.

5.2. Treatment of subordinate clauses

Marked formations and subordinate clauses, particularly relative clauses, also need more work:

(13) a. *Mi welodd y dyn y bachgen sy'n gweithio yn y siop.*
b. The man saw the boy that works in the shop.

(14) a. *Mi welodd y dyn y bachgen a giciodd y ci.*
b. *The man saw the boy and the dog kicked.
c. The man saw the boy who kicked the dog.

(15) a. *Hi yw'r ferch a welais ddoe.*
b. *She the daughter is and saw yesterday.
c. She is the girl I saw yesterday.

(16) a. *Ef yw'r dyn y lladdwyd ei fab.*
b. *He the man is were killed its son.
c. He is the man whose son was killed.

Part of this, as in (14) and (15), relates to improving the rules so that they make use of morphological markers where they exist (e.g. *a* (and) will not be followed by soft mutation, whereas *a* (who, which, that) will). Other work, as in (16), will involve trying to improve number and gender agreement.

5.3. Lexical selection

Lexical selection in Apertium is under development - this would increase fluency in examples such as the following:

(17) *big / great* (adjective)

a. *Mae hyn o bwysigrwydd mawr i Gymru ac yn hanfodol i ddyheadau'r Cynulliad.*

b. *This is of big importance to Wales and essential to the aspirations of the Assembly.

c. This is of great importance to Wales and esssential to the aspirations of the Assembly.

(18) *as / like* (conjunction)

a. *Mae'r adeilad hwn yn anaddas fel cartref hirdymor i'r Cynulliad.*

b. *This building is inappropriate like long-term home to the Assembly.

c. This building is unsuitable as a long-term home for the Assembly.

6. Discussion

The collaborative development model adopted from the free and open-source software movement has a great deal to offer MT research:

1. Existing data (e.g. dictionaries, corpora) can be reused, and software can be extended. Improvements can be shared, thus ensuring that scarce resources are used to best effect, and that the materials are as robust as possible by being tested in different contexts.

2. The robustness or versatility of a research model can be tested more rigorously. The fact that Apertium can be used for language families for which it was not designed, for instance, encourages us to believe that its architecture is broadly valid.

3. Results can be verified – the concept of reproducibility is central to the scientific method, but is absent unless all of the original data and software can be executed, examined and changed as necessary.

It must be emphasised, however, that in order for MT research to describe itself as "open" or "open-source" **both** the engines **and** the data (whether rules, dictionaries, grammars, corpora or whatever) need to be available under an open licence. For marginalised languages, the latter can sometimes be a problem, in that where such data exists it may not be open, due to an immature understanding on the part of language promotion bodies of the benefits of open research – see also (Streiter, Scannell, and Stuflesser, 2006, Forcada, 2006, Koster and Gradmann, 2004). Our strong view is that any materials developed with the input of public funds should by default have at least a subset released under an open licence. During the development of *apertium-cy* we have had reason to reflect on the odd situation of being able to use materials for English developed with public money in Spain, but not materials for Welsh developed with public money in Wales.

For SMT, given a corpus for any pair of languages, it is usually possible to arrive at a working MT system after a relatively short training time. In contrast, RBMT systems have often taken years to develop, largely because of the time involved in writing the dictionaries and the rules. We would argue that for marginalised languages the collaborative development model improves the attractiveness of RBMT as an MT option, for a number of reasons:

1. Such languages may not have freely available aligned corpora available of the size required for SMT,[13] but if they have dictionaries and grammars available, writing rules based on these will be as viable as trying to generate corpora.
2. Development time for open RBMT can be drastically reduced compared to closed RBMT by editing and adapting any open data that is already available in some form (e.g. dictionaries put together by enthusiasts). The authors are currently combining free Breton–Dutch and Breton–French dictionaries to create a unified 50,000-word resource, to which English can be added in due course.
3. If dictionaries and rules already exist for a series of language pairs, they can be edited and adapted to create a new pair. With Apertium, for instance, given the existing Spanish-English and Welsh-English pairs, the data and rules could be used to create a Welsh-Spanish MT system. Although the same point applies to SMT, translating a corpus may be a much more demanding affair.
4. For marginalised languages, the "critical mass" of required data or language tools may be no less than for a majority language, but the resources available (skills, time, money) are usually much less. Compared to the balkanisation of effort represented by closed systems, the decreased time (and therefore money!) required by open RBMT systems make it more feasible for such languages to get machine translation systems, and for this development to be done on a community basis.
5. Further useful tools may become available as a spin-off from the main effort – for instance, the first author was able to develop a proof-of-concept Welsh vocabulary assistant for web-pages, *Geriaoueg*, based on Apertium, in a couple of days.[14]

7. Conclusions

We have shown how a rule-based machine translation system for Welsh was quickly developed from existing data and software, and demonstrated that the translation quality in this initial version of the software is encouraging. Crucially, the materials used in the project were all available under a free licence which allowed them to be used and adapted in a collaborative development process. This also means that they are all available for review by interested researchers. We have pointed out a number of compelling reasons why the collaborative development model and rule-based MT systems are a good fit for marginalised languages.

Acknowledgements: We are grateful to Liam Tomkins and Dafydd Francis for their support during the development of the *apertium-cy* system, and to Felipe Sánchez Martínez, Gema Ramírez Sánchez, Sergio Ortiz Rojas and Mikel L. Forcada for their comments and suggestions.

[13] In fact, Welsh has a large corpus (the Proceedings of the National Assembly of Wales, referred to above) that could be used for SMT, but unfortunately it is not available under a free licence.

[14] `http://elx.dlsi.ua.es/geriaoueg`

Bibliography

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bello, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Miriam A. Scalco. 2006. "Open-source Portuguese-Spanish machine translation". *Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR-2006*.

Ball, Martin J. and Nicole Müller. 1992. *Mutation in Welsh*. Routledge.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. "Re-evaluating the role of Bleu in machine translation research". *EACL-2006*.

Crystal, David. 2000. *Language Death*. Cambridge University Press.

Forcada, Mikel. 2006. "Open-source machine translation: an opportunity for minor languages". *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages), LREC-2006*.

Garrido-Alenda, Alicia and Mikel L. Forcada. 2002. "Comparing nondeterministic and quasideterministic finite-state transducers built from morphological dictionaries". *Procesamiento del Lenguaje Natural (XVIII Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural)*.

Jones, Dafydd and Andreas Eisele. 2006. "Phrase-based statistical machine translation between English and Welsh". *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages), LREC-2006*.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila. 1995. *Constraint Grammar: A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Koster, Cornelis H. A. and Stefan Gradmann. 2004. "The Language belongs to the People". *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC-2004*.

Labaka, Gorka, Nicholas Stroppa, Andy Way, and Kepa Sarasola. 2007. "Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation". *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC-2004*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: A method for automatic evaluation of machine translation". *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Phillips, John D. 2001. "The Bible as a basis for machine translation". *Proceedings of PACLing 2001*.

Roche, Emmanuel and Yves Schabes. 1997. *Finite-State Language Processing*. MIT Press.

Scannell, Kevin P. 2008. "*An Gramadóir*: A grammar-checking framework for the Celtic languages and its applications". *14th annual NAACLT conference*.

Somers, Harold. 2004. *Machine Translation and Welsh: The Way Forward*. Welsh Language Board. Available from `http://www.byig-wlb.org.uk/english/publications/publications/2302.doc`.

Stewart, T.W. 2004. *Mutation as Morphology: Bases, Stems and Shapes in Scottish Gaelic*. Doctoral dissertation, Ohio State University.

Streiter, Oliver, Kevin P Scannell, and Mathias Stuflesser. 2006. "Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers". *Machine Translation*, 20(4).