Machine translation for closely related language pairs

Kevin P. Scannell

Department of Mathematics and Computer Science
Saint Louis University
St. Louis, Missouri, USA 63103
scannell@slu.edu

Abstract

We exploit the close linguistic relationship between Irish and Scottish Gaelic to develop a robust machine translation system `ga2gd`, despite the lack of full parsing technology or pre-existing bilingual lexical resources.

1. Introduction

Irish (Gaeilge) and Scottish Gaelic (Gàidhlig) are close linguistic relatives, forming, together with Manx Gaelic, the Goidelic (or Q-Celtic) branch of the Indo-European family. For simplicity in what follows, we will refer to the two languages simply as "Irish" and "Scottish" (in spite of the fact that the latter term, when used in isolation, generally refers to the variety of English spoken in Scotland).

Both languages are spoken daily by small minorities (measured in the tens of thousands), primarily within geographic strongholds situated in each case on the margins of an overwhelmingly English-speaking country. Despite enthusiastic communities of learners scattered throughout the world, the numbers of native speakers have continued to decline over the past century.

In §2, we discuss in some detail the extent of the linguistic similarity between Irish and Scottish. For now we simply point out that the the languages are not, generally speaking, mutually intelligible. They have distinct orthographies, independent (and diverging) lexica, and a number of important structural differences in terms of syntax.

Nevertheless, the languages are close enough that high-quality machine translation can be achieved with a manageable number of syntactic transfer rules, at least when combined with robust, statistically-based word sense disambiguation. Furthermore, by leveraging the existing open source language resources developed by the author, a complete system was implemented without unreasonable effort.

We believe there are a number of similar under-resourced language pairs that could benefit from a comparably naive approach, (e.g. Zulu ↔ Xhosa, Hiligaynon ↔ Cebuano, or Tokelauan ↔ Tuvaluan). Perhaps also of interest are pairs in which one language is a global one (e.g. French ↔ Walloon, or Italian ↔ Sardinian), where robust MT in either direction would be immensely useful. This has already been achieved, for example, by the open source project *Apertium* (Corbí-Bellot et al, 2005), which uses a design quite similar to ours for Spanish ↔ Catalan and Spanish ↔ Galician MT.

We owe great thanks to Caoimhín Ó Donnaíle for providing a tremendous amount of machine-readable data for Scottish, and for sharing his substantial linguistic expertise (in both languages).

2. Linguistic comparison

Since the robustness of the translator depends to a great extent upon the close linguistic relationship between Irish and Scottish, we thought it important to include in this section an indication of just how close this relationship is. For simplicity, and because we are interested primarily in translating electronic documents, we focus on the written languages. More details on this subject can be found in (Ó Rathaille, 1932, Ch. XVI), (Mac Maoláin, 1962), and (McCone, 1994).

The main difficulty with making such a comparison is the fact that the languages in question are "moving targets" in the sense that a given text (in, say, Irish) is parameterized in a number of ways that influence its relationship with Scottish: e.g. the date it was written, the regional dialect in which it was written, or its linguistic register.

All three Goidelic languages share a common ancestor in *Middle Irish*, forms of which were spoken in Ireland, on the Isle of Man, and in the Scottish Highlands until roughly the 16th century. Even after the spoken languages had diverged, there was a shared literary tradition written in the so-called *Gaeilge Chlasaiceach* (Classical Gaelic) up through the 18th century; this was the language of most of the early printed books in both languages: from Carswell's translation of Knox's liturgy in 1567 to the Bible translations of the 17th century. Note, however, that by 1690 the languages had diverged to the extent that when the Irish translation of the Bible was reprinted in Roman characters for the benefit of Gaelic speakers in Scotland, there were complaints that the text was unreadable (Williams, 1986, pp. 101–102).

Geographically, there was at one time a continuum of dialects ranging from the far southwest of Ireland to the northernmost parts of Scotland. As a consequence, the Ulster dialect of Irish, spoken in northeastern Ireland, is by far the closest to Scottish. We note, however, that in practice these differences are of little importance to the translator: input texts are normalized in various ways that minimize any dialect differences that might be present; see §3.2.

The single greatest disaster in terms of mutual understanding between the languages was the introduction of the *Caighdeán Oifigiúil* (Official Standard) on the Irish side in the 1940's (Rannóg an Aistriúcháin, 1962). As an example, the Scottish Gaelic words *bàgh* (bay), *bàidh* (sym-

pathy), and *bàthadh* (drowning) are immediately recognizable and distinguishable in pre-standard Irish (Dinneen, 1927) as *bádh*, *báidh*, and *bádhadh*, respectively, while the *Caighdeán* tragically conflates all three into the indescript "*bá*" (Ó Dónaill, 1977). Similar examples abound.

There was also a spelling reform on the Scottish side, put forward in 1981 with the publication of the "Gaelic Orthographic Conventions" (GOC) document[1]. These reforms were less sweeping, and offered more of a mixed bag in terms of the relationship with Irish. On the one hand, the GOC changed things like the shared acute accents on words like *mór* or *féin* to grave accents, but at the same time made other words look more Irish, by mimicking some of the *Caighdeán* reforms (e.g. replacing *sg* with *sc* and *sd* with *st*).

For the benefit of readers completely unfamiliar with these languages, we will attempt to make the preceding discussion a bit more concrete by offering a single sentence[2] rendered in each language; we will draw upon this example occasionally in the sections below.

> Cén fáth a bhfeiceann tú an cáithnín i súil do bhráthar agus nach n-airíonn tú an tsail i do shúil féin?
>
> Agus c'ar son a tha thu a' faicinn an smùirnein a tha 'an sùil do bhràthar, ach nach 'eil thu 'toirt fainear na sail' a tha ann do shùil féin?

3. Design and implementation

3.1. Overview

The `ga2gd` software is implemented, from the perspective of an end-user, as a standard Unix filter:

```
$ echo "lá breá éigin" | ga2gd
latha brèagha air choireigin
```

The same is true of the internal architecture; an input text is piped through a sequence of smaller standalone components which transform the text in various ways:

1. Irish standardization.
2. POS tagging, stemming, and chunking.
3. Word sense disambiguation.
4. Syntactic transfer.
5. Lexical transfer.
6. Scottish post-processing.

The sections below describe each of these components in brief, and, where appropriate, some indication is given of how they were assembled.

3.2. Irish standardizer

In recent work, the author created a web crawler and search engine tailored specifically to the Irish language documents on the web[3]. In contrast with general-purpose search engines, which tokenize and index the documents they find in exactly the form in which they appear on the web, our site also converts Irish documents to a form approximating the *Caighdeán Oifigiúil* and indexes them both ways. This allows access to all historical (or dialect) Irish documents on the web through a single, simple search interface.

The standardizer amounts to a finite state transducer that encodes the morphological rules of non-standard Irish together with mappings to standardized forms. These rules are augmented with a large database of non-standard/standard word pairs that was extracted in part from a parallel corpus of English and Irish texts (Scannell, 2005).

This phase is important in that it allows `ga2gd` to translate non-standard Irish as easily as standard Irish. It also has the advantage that, when constructing the bilingual lexicon, one need only provide Scottish translations for standard citation forms of Irish words.

3.3. Irish tagger, stemmer, and chunker

There is no full-scale parsing technology available for either Irish or Scottish, and, consequently, there are no treebanks one could use to implement a statistical MT system. There are, however, robust (rule-based) part-of-speech taggers built into the open source Gramadóir grammar checker[4] for each language. Irish, in addition, has a rule-based *chunker*, which delimits chunks in the spirit of (Ramshaw and Marcus, 1995). As it turns out, the syntactic differences between the languages are small enough that chunking is sufficient (in most cases) for finding accurate translations; this is discussed below in §3.5.

When this phase is completed, XML tags have been added to each word in the input stream that indicate the word's part of speech, its stem, and the stem's part of speech. For example, in our example sentence, *bhfeiceann* is transformed into:

```
<w>
 <t>
  <V p="y" t="láith">bhfeiceann</V>
 </t>
 <s>
  <V p="y" t="ord">feic</V>
 </s>
</w>
```

The stems and their POS tags are used in an essential way by the word sense disambiguation module; see the next subsection.

This component of the pipeline coincides almost exactly with the standalone Gramadóir grammar checker for Irish, and so we direct the reader to that project's documentation for further implementation details[5].

3.4. Irish word sense disambiguation

Because of the syntactic similarities between the languages, it turns out that the stickiest translation problems are, for the

[1] See `http://www.smo.uhi.ac.uk/gaidhlig/goc/`.

[2] Matthew 7:3: "Why do you see the speck that is in your brother's eye, but don't consider the beam that is in your own eye?"

[3] See `http://www.aimsigh.com/`.

[4] See `http://borel.slu.edu/gramadoir/`.

[5] Ibid.

most part, semantic. Solving these problems relies upon a robust word sense disambiguation (WSD) system.

For `ga2gd`, the WSD filter is implemented as a naive Bayes classifier. It takes as input a tagged and chunked Irish text, and begins by searching in each sentence for words that have more than one possible Scottish translation. When such a word is found, a *feature vector* is generated which is made up of the stemmed and tagged words from the sentence, plus features indicating whether or not the words adjacent to the ambiguous word have initial mutations[6]. Then the most probable sense for the ambiguous word is chosen, given the computed feature vector, and this sense is added to the text stream as an attribute within the `<t>` tag from the previous subsection, e.g. `<t sense='1'>`. The appropriate probabilities are computed using training data bootstrapped from a small, manually disambiguated corpus.

It is critical in a number of instances to consider the mutations on adjacent words. For example, the Irish adjective *céad* can mean "first" or "one hundred" and precedes the noun it modifies in each case. When it means "first", however, it causes lenition of the modified noun: *a céad cheacht* "her first lesson", but *céad bliain ó shin* "a hundred years ago". Without this clue, there is very little else (statistically speaking) that one might rely upon to perform this disambiguation.

The *bá* example mentioned above is a good one, though note that the part-of-speech tagger shares the responsibility for distinguishing the masculine "drowning" sense from the other (feminine) senses. A similar example is Ir. *fiach*, which can mean "a debt, obligation", "a raven", or "a hunt", and these senses are translated to Scottish as *fiach*, *fitheach*, *fiadhach*, respectively[7].

Ambiguous words are quite common. Looking up a random sample of words from our database in (Ó Dónaill, 1977) indicates that between 10% and 20% of words have multiple senses[8]. At present, the WSD module has been trained for less than 1000 ambiguous Irish words, though we hope to grow this to about 3000, or approximately 10% of the total lexicon. This should be more than adequate for accurate translation to Scottish since not infrequently an ambiguity in Irish is shared on the Scottish side, e.g. *bonn* can mean either "base, foundation" or "coin, medal" in both languages. This is another important way in which translation between closely related languages is substantially easier than the general case.

[6] An *initial mutation* in Irish or Scottish (or the other Celtic languages) is a phonological change that occurs at the beginning of a word in certain situations, usually depending on the syntactic relationship with, or some grammatical feature of, the preceding word. An important example in both languages is *lenition*, which is indicated orthographically by the insertion of an "h" after an initial consonant. Irish has another consonant mutation called *eclipsis* that has no orthographic counterpart in Scottish.

[7] And note that, as with *bá*, the *Caighdeán Oifi gŭil* is partially responsible for the ambiguity in this case: the "hunting" sense is generally spelled *fi adhach* in pre-standard Irish

[8] This percentage is probably larger than what one would get by sampling the dictionary in full; our database omits a large number of (generally monosemous) rare or technical terms.

Even our simple example sentence from the end of §2 requires disambiguation of at least three words: *cáithnín*, *súil*, and *(t)sail*. The word *súil* illustrates a recurring issue for the language pair in question. Here it means "eye" but also functions as a so-called "verbal noun" after *ag*, e.g. *ag súil le* "hoping/waiting for", and the cognate translation *sùil* is not acceptable in the latter case. The same situation arises with many other Irish words such as *cnuasach*, *cruinniú*, *scrúdú*, etc.

In the sample sentence, the word *sail* means "beam, stick", but quite commonly means "dirt, dross" in the Irish corpus, and these senses translate to distinct Scottish words (*sail* (feminine) and *sal* (masculine) respectively). Similarly, the word *cáithnín* means something like English "speck, mote" in the present example, but in theoretical physics is used for things like subatomic particles. The WSD module has no difficulty distinguishing the latter sense since it often appears in sentences together with unambiguously technical terms such as *cosmach* "cosmic", *treoluas* "velocity", or *déacht* "duality" (though this is an instance in which disambiguation is less important – the single Scottish term *smùirnean* is probably a safe translation in either case).

3.5. Context-sensitive syntactic rewriting

As noted above, we do not have a full parse of the input sentence at this stage, but instead something resembling a parse tree of depth one.

An important example of a syntactic rewrite rule comes from the fact that there is no exact analogue of the present tense Irish verb in Scottish. This is why the phrase *(bh)feiceann tú* "you see" is translated as *tha thu a' faicinn* "you are a'seeing" in our example above. In cases like this, once the chunker has correctly identified the subject noun phrase of the present tense verb, then a simple syntactic transfer is sufficient:

(S (VBZ x) (NP y)) → tha t[y] a' t[x]

where the mapping $x \rightarrow t[x]$ means to recursively translate the given constituent, and in the special case of present tense Irish verbs we map to the appropriate verbal noun in Scottish.

The Irish imperfect tense is also not available in Scottish, and so we have the following similar rewrite rule:

(S (VBI x) (NP y)) → b'àbhaist do t[y] a bhith a' t[x]

The transfer rules are stored in a plain text input file and are expressed in a syntax similar to the examples above. Then, before the actual translation process begins, each rule is transformed into a finite state recognizer which can be compiled for exceptionally fast matching against the (tagged and chunked) input stream.

The current version has just under 100 transfer rules, though we expect this number to grow rapidly as we continue to add rules for handling additional multi-word phrases.

3.6. Bilingual lexicon

A number of different techniques were used to construct the Irish-Scottish lexicon required by the translator. Because of the scarcity of parallel texts in the two languages, we were unable to exploit mutual information techniques to any great extent (though a small number of word pairs were extracted by aligning the electronic Bible texts in the two languages).

At least 90% of the translations in the lexicon were extracted automatically from two existing electronic dictionaries, one Irish-English and one Scottish-English. The first of these was constructed by the present author while constructing a monolingual Irish thesaurus (Scannell, 2003). The Scottish-English data were provided by Caoimhín Ó Donnaíle, from among the many resources he manages at Sabhal Mòr Ostaig[9].

It was deemed desirable in constructing the bilingual lexicon to select cognate translations when possible, even at the risk of making the Scottish translations sound a bit "Irish", as a way of emphasizing (and maybe reinforcing) the common literary heritage of the two languages.

Finding cognates is straightforward. We first used a *fine-grained mode*, which applies a number of simple spelling changes to Scottish words to make them look as "Irish" as possible (grave accents made acute, *chd$ → cht$*, *achadh$ → ú$*, *sg → sc*, etc.). Pairs are then deemed to be cognates if the normalized Scottish word has edit distance zero or one from the Irish word, and if, in addition, they share at least one English translation. The *coarse mode* works similarly, but in this case both Irish and Scottish words are converted to a coarse phonetic encoding (originally used to implement Philips' metaphone algorithm in our Irish `aspell` spellchecker[10]); this approach generated pairs of cognates with fewer false positives than by merely increasing the allowable edit distance in the fine-grained mode.

The requirement that potential cognates share at least one English translation was sufficient to avoid all examples of *faux amis* known to us. In some instances we were saved by the limited size of the Irish-English and Scottish-English databases. For example, there were no English translations in common for *cuan* (Ir. "bay, harbor, port, inlet, haven", Sc. "ocean, sea"), though the "harbor" sense appears (with the qualification "rarely") in Dwelly's magnificent unabridged Scottish dictionary (Dwelly, 2001).

Finally, when no cognates were found for a given Irish word, a "best guess" translation was made automatically using a metric based on the number of shared English translations and the corpus frequencies of the Scottish words. In all, we were able to produce translations for 21,106 Irish lemmas with this approach; 8462 of these have been verified by hand against print dictionaries, and evaluated for potential disambiguation.

[9] This is a good illustration of the power of refactoring existing resources for minority languages (even resources designed and built with entirely different projects in mind), and argues for making all such data freely available under an open source license.

[10] See `http://aspell.net/metaphone/` and `http://borel.slu.edu/ispell/index-en.html` for more information.

Note that the lexicon only pairs up Irish citation forms with their Scottish equivalents. A morphological generator for each language is then employed to pair up the corresponding inflected and mutated forms – at present this amounts to nearly 200,000 distinct Irish words that the translator can handle.

We have not as yet made any attempt at evaluating precision, but we have some preliminary results on the (word-for-word) recall of the translator. First of all, we should point out that our aim is to have the translator handle a very broad range of texts from different genres and literary registers (newspaper articles, government documents, email lists, newsgroups, blogs, etc.) since it will eventually be applied to the full web corpus of Irish (see §4). With this in mind, we performed an evaluation on a corpus of 1.89 million words of text crawled from the web. As a baseline, note that the Gramadóir spelling and grammar checker underlying `ga2gd` (and sharing the same Irish lexicon) recognized 91.14% of the words in the corpus (the others consisting mostly of English pollution, but also some misspellings and a few truly unrecognized words). The recall of the translator, as measured on the subset of 1.72 million known words, was 92.72%.

3.7. Scottish Post-processing

To this point we have not thought very carefully about generating grammatically correct Scottish sentences. For this, we use the nascent Scottish version of the Gramadóir grammar checker to automatically make any necessary local corrections when there are incorrect initial mutations, etc., in the naively generated output.

We have used a similar approach in English → Irish MT, where one can blithely translate a fragment like "the man" as "an fear" without considering the wider context, but then in post-processing, if the wider context happens to be "with the man", one obtains "le an fear" which is corrected to "leis an bhfear" by the grammar checker. This approach allows the complexity of the translator to be focused where it belongs (on global syntactic issues and WSD), and moves trivial post-processing to an independently useful (and independently developed) monolingual application.

4. Applications: Cross-language Information Retrieval

The `ga2gd` software will be integrated into the Irish language search engine mentioned above as a "Translate this page" feature, allowing Scottish speakers to browse and read the substantial amount of Irish language material on the web. The eventual aim is to combine all three Goidelic languages in a single search engine, where queries can be made in one language, with all relevant documents in any of the others being returned, and translated if necessary.

5. References

Patrick S. Dinneen, editor. 1927. *Foclóir Gaedhilge agus Béarla*. Irish Texts Society, Baile Átha Cliath.

Edward Dwelly, editor. 2001/1912. *Illustrated Gaelic-English Dictionary*. Birlinn Ltd., Dùn Èideann.

Seán Mac Maoláin. 1962. *Gàidhlig agus Gaeilge*. An Gúm, Baile Átha Cliath.

Kim McCone, editor. 1994. *Stair na Gaeilge*. Roinn na Sean-Ghaeilge, Coláiste Phádraig, Maigh Nuad.

Niall Ó Dónaill, editor. 1977. *Foclóir Gaeilge-Béarla*. An Gúm, Baile Átha Cliath.

Tomás Ó Rathaille. 1932. *Irish dialects past and present*. Institiúid Árd-Léinn, Baile Átha Cliath.

Antonio M. Corbí-Bellot, Mikel L. Forcada, et al. 2005. An open-source shallow-transfer machine translation engine for the Romance languages of Spain In *Proceedings of the 10th Annual EAMT Conference*, Budapest, Hungary, 30-31 May 2005.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In D. Yarovsky and K. W. Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94. Association for Computational Linguistics, Somerset, New Jersey.

Rannóg an Aistriúcháin. 1962. *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil*. Oifig an tSoláthair, Baile Átha Cliath.

Kevin P. Scannell. 2003. Automatic thesaurus generation for minority languages: an Irish example. In *Actes de la 10e conférence TALN à Batz-sur-Mer*, volume 2, pages 203–212. ATALA.

Kevin P. Scannell. 2005. Applications of parallel corpora to the development of monolingual language technologies. Preprint.

Nicholas Williams. 1986. *I bPrionta i Leabhar*. An Clóchomhar, Baile Átha Cliath.