Proceedings of the First Celtic Language Technology Workshop, pages 33 40,
Dublin, Ireland, August 23 2014.

(tens of millions of words) produced between about 1925 and 1945 is effectively unusable for language modeling or other NLP applications.

This explains the importance of Irish standardization, but what does it have to do with Scottish Gaelic–Irish MT? The answer is twofold: first, we can cast the standardization problem as an MT problem between two *very* closely-related languages (namely, pre-standard and standard Irish), and second, the orthography of pre-standard Irish has a great deal in common with the orthography of Scottish Gaelic, and so it turns out that a single statistical model works well to solve both problems. A brief description of the standardizer has appeared previously in (Uí Dhonnchadha et al., 2014), as part of a more complex pipeline for processing historical Irish texts for lexicography.

Many authors have considered MT between closely-related language pairs, including dozens of papers describing rule-based systems based on the Apertium engine (Forcada et al., 2011).[2] Several other papers have taken statistical or hybrid approaches, e.g. (Hajič, 2000) for Czech and Slovak, (Altintas and Cicekli, 2002) for Turkish and Crimean Tatar, (Nakov and Tiedemann, 2012) for Macedonian and Bulgarian, and (Miller, 2008) and the references therein. Statistical MT techniques have also been used previously for historical text normalization; see for example (Pettersson et al., 2013).

The outline of the paper is as follows: In section 2, we describe the shared statistical model in general terms, and discuss the problematic notion of "standard Irish". In section 3 we describe `gd2ga`, our Scottish Gaelic–Irish MT system, along with the parallel corpus used in its development. Finally, in section 4, we introduce and evaluate the Irish standardizer.

2 The Model

In this section we describe the statistical model that underlies both the `gd2ga` machine translation system and the Irish standardizer. We view both problems as instances of machine translation between very closely-related languages, the latter requiring translation from what we will call "pre-standard Irish" to "standard Irish" (with scare quotes because both terms are problematic; more on this below). Because of the limited syntactic differences between source and target in each case, it suffices to use a simple word-based model without reordering, a variant of the well-known IBM model 1 (Brown et al., 1993). It is important to distinguish the statistical model per se (which assigns probabilities to translation candidates) from the means by which those probabilities are acquired. In the context of the IBM models, Expectation Maximization (EM) is typically used for the latter; here we take a simpler approach, described in section 2.2. In the two subsections that follow, we will describe the language model and translation model, respectively.

2.1 Language Model

The target language in both cases is what we are calling "standard Irish", and so a single language model suffices for both systems. We use a trigram model which is typical in this context and normally would require little further comment. In our situation, however, we run into a couple of major difficulties.

First, there is not complete agreement with respect to what "standard Irish" means. The first movement toward standardization of the written language goes back to the 1930's with the establishment of government committees to look at the question. A simplified spelling system was published in 1945 (Rannóg an Aistriúcháin, 1945) and implemented by the government translation office around the same time. A simplified grammar was published in 1958 (Rannóg an Aistriúcháin, 1958), followed by two important bilingual dictionaries (de Bhaldraithe, 1959; Ó Dónaill, 1977) that helped encourage use of the standard language by the general public. The problem is that these dictionaries do not completely conform to the published standard, nor do many grammar books that are used in schools and by independent learners even today. To compond the confusion, a revised version of the official standard was recently published (Uíbh Eachach, 2012), and has been criticized by some in the language community (Mac Lochlainn, 2012), so it remains to be seen the extent to which it will be embraced as the new "standard Irish".

The second difficulty is that, even to the extent that everyone agrees on certain elements of the standard, *no one* implements them completely in their writing, which is to say that virtually all non-trivial texts in

[2]For the most up-to-date references, see `http://wiki.apertium.org/wiki/Publications`.

Irish contain some non-standard forms. In the case of `gd2ga`, this is not of great concern; we could train the target language model with the texts we have, and the output would resemble the fluent, natural Irish of the training texts. For the standardizer, however, we are aiming at very high precision, with the output conforming to *some* version of the standard. In short, the problem is this: we want to train an n-gram model for a certain language, but there are *no non-trivial texts written in that language*.

We get around these issues by making use of a suite of open-source Irish language proofing tools called *An Gramadóir*.[3] From an initial corpus of about 100 million words, we selected a subset of about 40 million words comprised of the texts that are most conformant to the standard. To do this, the rules implemented in *An Gramadóir* were first separated into those representing true "errors" (misspellings, grammatical mistakes, etc.) from those representing standardizations. Then, *An Gramadóir* was run on every text in the corpus in order to assign each a numerical measure of "non-standardness" (the number of standardizations flagged per 100 words). The subcorpus was chosen from the texts with the lowest non-standardness scores. Finally, we applied a small number of automated substitutions to certain non-standard forms that are nearly as common as their standardizations in real-world texts, e.g. *nach dtáinig* vs. *nár tháinig* ("did not come").

Once the training corpus is generated in this way, we tokenize and compute the probabilities for the trigram language model (no pruning), and smooth using absolute discounting (Chen and Goodman, 1996). The implementation of the language model is included as part of the translator itself in order to avoid external dependencies on libraries such as IRSTLM, KenLM, etc.

2.2 Translation Model

The translation model represents the conditional probability of some source language word corresponding to a given target language word. Since the parallel corpora for our two translation problems are relatively small, and since our goal is very high-precision translation, a statistical word alignment approach using Expectation Maximization does not give suitable results. Instead, we take advantage of the resources that we have at hand; specifically, high-quality bilingual lexicons together with a well-understood set of spelling rules for mapping source language words to cognates in the target language. In the context of Scottish Gaelic to Irish MT, the latter include mappings like *-chd* → *-cht* and *-eu-* → *-éa-* (together these two rules map a word like *creuchd* ("wound") to its Irish cognate *créacht* for example). For Irish standardization, there is a separate set of rules but with significant overlap, e.g. *-idhea-* → *-ío-*, which maps *buidheach* ("thankful") to *buíoch* and *fuidheall* ("remainder") to *fuíoll*. These last two examples are valid for both `gd2ga` and for the standardizer.

A source-to-target language mapping is often discovered through a combination of rule-based spelling changes like the ones above, plus a lexical mapping when the rules do not suffice. We define the translation model in the following simple way: (1) all source language words that are paired with a given target language word are assumed to have the same conditional probability; and (2) when a source language word is paired with a target language word by applying some number n of spelling rules, we multiply the conditional probability by a fixed "penalty" β^n. When more than one sequence of rules leads to the same target word, we take the shortest sequence. The constant β was optimized through a tuning process using held-out data from the parallel corpora; we used the value of β which gave the smallest word-error rate on the held-out test set.

Choices (1) and (2) above were made for the sake of simplicity. In future work, we plan to experiment with allowing different probabilities for different rules, as well as using EM to train the translation probabilities, restricting to the lexicographical translation pairs.

The decoding process is quite simple. The algorithm processes the source sentence word-for-word from left-to-right, and keeps track of all possible target language hypotheses along with their probabilities, as computed using the translation model and language model. When multiple hypotheses share the same final two words, we are able to discard all but the one with the highest probability. When we reach the end of the sentence, the highest probability candidate is output as the translation.

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

3 Scottish Gaelic to Irish MT

Scottish Gaelic and Irish are are sufficiently distinct as spoken languages that even fluent speakers without experience in the other language are usually only able to understand bits and pieces. As a result, there are very few spoken-language contexts in which Scottish Gaelic and Irish speakers are able to interact with each other in either language, and often have to resort to English.[4]

The written language is significantly easier, even in light of the Irish spelling reform and more recent reforms on the Scottish Gaelic side (Scottish Qualifications Authority, 2009) which have made things more difficult than they might conceivably be. Indeed, there are vibrant online communities of Irish and Scottish Gaelic speakers availing themselves of social media, especially Facebook and Twitter, and there is evidence of a significant amount of bilingual communication going on between the two language communities. (Scannell, 2013)

We believe there could be even more, given the right tools. By implementing high-quality Scottish Gaelic to Irish machine translation, and by deploying it in combination with our earlier *ga2gd* system, we hope to encourage greater communication between the two communities.

3.1 Parallel Corpus

A parallel corpus plays a key role in the development of the bilingual lexicon and spelling rules, as well as being used for evaluation purposes. Unfortunately, direct translations between the two languages are extremely rare (despite the relative ease with which such translations could be made), and even translations of a common English source text proved hard to come by. So we chose to include quite a bit of material that might otherwise have been left out of a parallel corpus: software translations (Firefox, LibreOffice, etc.), poetry, song lyrics, prayers, bilingual word lists, Irish glosses on Scottish Gaelic source material (and vice versa), bilingual tweets, titles of linked Wikipedia pages, and so on. When combined with more traditional material (Bible texts, fiction and non-fiction prose translations), we were able to assemble roughly a million words of parallel text: 129,983 translation segments containing 1,016,041 words of Scottish Gaelic and 956,598 words of Irish. This is, to our knowledge, the only non-trivial parallel corpus for this language pair.[5]

3.2 Bilingual Lexicography

The heart of the system is the bilingual lexicon which is being painstakingly constructed by hand (work in progress), drawing upon a number of freely available resources for both languages. Even though the translation system does no part-of-speech tagging, the lexicon stores lemmas in Scottish Gaelic tagged by part-of-speech, mapped to lemmas in Irish, also tagged by part-of-speech. Then, mappings between surface forms are produced by employing morphological generators on both sides (cf. (Tyers, 2009)). This produces mappings for over 150,000 surface forms from a bilingual lexicon with about 13,000 lemmas.

We have used the following resources while building the lexicon.

- The parallel corpus described in section 3.1
- Scottish Gaelic–English dictionaries created by Michael Bauer (Bauer, 2014)
- Various Scottish Gaelic–English dictionaries hosted by Sabhal Mòr Ostaig[6]
- The bilingual lexicon created for (Scannell, 2006)

[4] This despite the efforts of organizations like *Colmcille* (formerly *Iomairt Cholm Cille*), established to encourage precisely this sort of interaction.

[5] We have made the portions of the corpus that are under open licenses available here `http://borel.slu.edu/pub/ccgg.zip`.

[6] See `http://www2.smo.uhi.ac.uk/gaidhlig/faclair/`.

3.3 Evaluation

We began by evaluating the coverage of the source language lexicon. For this, we gathered a monolingual Scottish Gaelic corpus comprised of 3.9M tokens from 14713 web-crawled texts (Scannell, 2007). The system recognized 96.74% of the tokens in this corpus, a result which is comparable to, or even better than, the coverage of many open-source spell checkers on (noisy) web texts. This result is due to (1) the fact that we were able to re-purpose a number of open-source lexical resources when building our dictionary, (2) the addition of a large database of "untranslatables": proper names (e.g. *Facebook*, *Akerbeltz*), non-Gaelic words (mostly English, but some Latin, French, etc.), and abbreviations (e.g. *km*, *vs*) and (3) the ability of the system to handle misspellings and variants either by including them in the lexicon (with mappings to "standard" forms) or through the application of spelling rules.

Evaluating the MT system itself proved problematic. Even though we were able to assemble a parallel corpus, the vast majority of the texts are translations from a common English source (principally, the open-source software translations and the Bible texts), as opposed to direct translations between Irish and Scottish Gaelic. To get around this issue, the author manually translated a collection of 593 sentences from Scottish Gaelic to Irish and used this as a test corpus. When comparing the output of `gd2ga` with these reference translations, the word-error rate (WER) was 38.67%. This can be compared with a baseline system that simply leaves the Scottish Gaelic source text unchanged, yielding a WER of 88.09%.

This is still not completely satisfactory as an evaluation for a couple of reasons. First, given the nature of the statistical model, the translations produced by `gd2ga` stay very close to the source language text, and so a sentence like

> *Tha mi a' tuigsinn a-nis.*
> ("I understand now.")

translates to

> *Tá mé ag tuiscint anois.*

whereas a human translator might render this more naturally in Irish as "*Tuigim anois.*". Similar examples in other verb tenses abound. Secondly, the system sometimes gets initial mutations wrong (tending to be conservative and preserving the mutations of the source text due to the penalty factor β), though this rarely impacts comprehension or fidelity of the translation. It might be reasonable to compute a modified WER for Celtic languages that ignores differences in mutations, but we did not pursue this.

4 Irish Standardizer

The standardizer described in this section takes as input an Irish language text and outputs a version that conforms as closely as possible to "standard Irish", subject to the vagaries discussed above in section 2.1. The principal application of the standardizer has been to the indexing of pre-standard texts to enable search and retrieval via standard spellings, mainly for lexicographical purposes (Uí Dhonnchadha et al., 2014). In this application, the standardized texts are used only for indexing purposes, which is to say that the pre-standard texts are displayed to users in search results.

An interesting second application would be to apply the standardizer to the huge amount of Irish language literature (novels, plays, many of them in translation) published from the 1920's through the 1940's in order to make those texts accessible to a modern readership that has grown up on the standard orthography. Indeed, a number of these books have been republished in recent years, but to my knowledge they have all been standardized by hand, e.g. (Doyle, 2012). To do this automatically, somewhat more care would be needed in order to not completely destroy the richness of the Irish dialects found in these texts (as the standardizer in its current form does, more or less), probably by creating customized versions of the standardizer for each dialect, together with limited post-editing.

4.1 Parallel Corpus

To support development of the "bilingual" lexicon (pre-standard/standard word pairs) and spelling rules, and to enable formal evaluation of the system, we created a large parallel corpus of pre-standard/standard

sentence pairs. The standardizations were taken from republications of older material of the kind described above, and were performed manually by human editors. In all, we used eighteen books together with their standardizations, segmented by sentences and aligned into 46,914 translation pairs (almost all one sentence to one sentence). There are 814,365 words on the pre-standard side and 801,236 words on the standard side.

4.2 Lexicography

The bilingual lexicon is similar in structure to the Scottish Gaelic–Irish lexicon described above in section 3.2, with pre-standard lemmas being mapped to standard lemmas, and morphological generators applied to each side to create mappings of surface forms. The lexicon again draws upon existing resources; first and foremost, about 22,000 standardization pairs used by *An Gramadóir* for spelling and grammar correction, along with an additional 10,000 pairs drawn directly from an electronic version of (Ó Dónaill, 1977). After applying the morphological generators, we end up with mappings for about 135,000 surface forms. Keep in mind, however, that the spelling rules play a more important role for the standardizer than they do for the Scottish Gaelic translator, and so the actual source language coverage on pre-standard Irish texts is significantly better than the number 135,000 might suggest.

4.3 Evaluation

We performed two evaluations of the standardizer.

The first evaluation is similar to the one we performed on `gd2ga`, described above in section 3.3. Namely, we held out a sample of 200 sentence pairs from the parallel corpus, applied the standardizer to the pre-standard sentences, and compared the results with the reference standardizations, yielding a WER of 9.86%. Of course the translation problem here is much easier, as illustrated by a baseline WER of 27.28% obtained by leaving the pre-standard texts unchanged.

System	WER	Baseline
`gd2ga`	38.67	88.09
Standardizer	9.86	27.28

Table 1: Summary of results (Word Error Rates)

As a second evaluation, we looked at *sentence-level* accurancy. The point here is that in most cases there really is one "optimal" standardization of a given input sentence and that should be our aim. For example, the pre-standard

> *Acht go h-ádhmhail bhí lucht síothchána thall agus i bhfos.*
> ("But, luckily, there were peaceful people on both sides.")

must, in a just world, map to:

> *Ach go hádhúil bhí lucht síochána thall agus abhus.*

and we would consider any other standardization as incorrect.

The second evaluation, therefore, involves holding out a sample of 4147 sentence pairs from the parallel corpus, applying the standardizer to the pre-standard sentences, and comparing word-for-word with the standardized sentence (ignoring differences in punctuation). The current version gets 35.06% of these sentences completely correct. This can be compared with a score of 7.45% for a baseline system that does nothing to the input text (that is, 7.45% of the pre-standard sentences require no standardization at all, mostly very short sentences).

Acknowledgements

First and foremost I would like to thank Michael Bauer for making all of his Scottish Gaelic resources freely available, and for his outstanding work translating open-source software packages into Scottish

Gaelic. When combined with my own translations into Irish, these make up fully half of the Irish–Scottish Gaelic parallel corpus. Thanks also to Caoimhín Ó Donnaíle for providing access to a wealth of Scottish Gaelic lexicographical material and for answering many linguistic questions over the years. Donncha King kindly provided copies of several hard-to-find Irish translations of Scottish Gaelic short stories, along with the originals. I am grateful also to Ciarán Ó Duibhín for a number of useful email conversations regarding Irish standardization. The teams from the New English-Irish dictionary project and the Royal Irish Academy's *Foclóir na Nua-Ghaeilge*, especially Pádraig Ó Mianáin, Cathal Convery, Ruairí Ó hUiginn, and Máire Nic Mhaoláin, manually corrected the output of early versions of the standardizer which greatly sped up development. Finally, thanks to Elaine Uí Dhonnchadha and Brian Ó Raghallaigh for bravely installing and running unwieldly versions of the standardizer. This work was partially supported by US NSF grant 1159174.

References

Kemal Altintas and Ilyas Cicekli. 2002. A Machine Translation System Between a Pair of Closely Related Languages. In *Proceedings of the International Symposium on Computer and Information Sciences*, 192–196.

Michael Bauer. 2014. *Am Faclair Beag: Faclair Gáidhlig is Beurla le Dwelly 'na bhroinn [An English–Scottish Gaelic dictionary incorporating Dwelly]* `http://www.faclair.com/`. Retrieved June 26, 2014.

Michael Bauer. 2014. Speech and Language Technology on a Shoestring and how to get there in a hurry. `http://www.akerbeltz.org/iGaidhlig/wp-content/uploads/2014/07/SALT-on-a-Shoestring.pdf`. Retrieved July 3, 2014.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *ACL '96 Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 310–318.

Antonio M. Corbí-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, and Kepa Sarasola. 2005. An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. In *Proceedings of the 10th European Associtation for Machine Translation Conference*, 79–86.

Tomás de Bhaldraithe. 1959. *English–Irish Dictionary*. Oifig an tSoláthair, Baile Átha Cliath.

Niall Ó Dónaill. 1977. *Foclóir Gaeilge-Béarla [Irish–English Dictionary]*. Oifig an tSoláthair, Baile Átha Cliath.

Elaine Uí Dhonnchadha, Kevin Scannell, Ruairí Ó hUiginn, Eilís Ní Mhearraí, Máire Nic Mhaoláin, Brian Ó Raghallaigh, Gregory Toner, Séamus Mac Mathúna, Déirdre D'Auria, Eithne Ní Ghallchobhair, and Niall O'Leary. 2014. Corpas na Gaeilge (1882-1926): Integrating Historical and Modern Irish Texts. *Proceedings of the Workshop "Language resources and technologies for processing and linking historical documents and archives" at LREC 2014*.

Arthur Conan Doyle. 2012. *Cú na mBaskerville [The Hound of the Baskervilles]*. Translated by Nioclás Tóibín, standardized by Aibhistín Ó Duibh. Evertype, Co. Mhaigh Eo.

Vivian Uíbh Eachach, ed. 2012. *Gramadach na Gaeilge: An Caighdeán Oifigiúil, Caighdeán Athbhreithnithe [Irish Grammar: The Official Standard, Revised Standard]*. `http://www.oireachtas.ie/parliament/media/Final-Version.pdf`. Seirbhísí Thithe an Oireachtais, Baile Átha Cliath.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation *Machine Translation*, 25(2):127–144.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *ANLC '00 Proceedings of the sixth conference on Applied natural language processing*, 7–12.

John Judge, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Elaine Uí Dhonnchadha, and Kevin Scannell. 2012. *An Ghaeilge sa Ré Dhigiteach [The Irish Language in the Digital Age]*. Springer-Verlag, Berlin.

Antain Mac Lochlainn. 2012. Léirmheas ar an Chaighdeán Oifigiúil, 2012 [Review of the Official Standard, 2012]. http://www.aistear.ie/news-details.php?ID=33. Retrieved May 2, 2014.

Bryce Miller. 2008. Translating Between Closely Related Languages in Statistical Machine Translation. MS Thesis, University of Edinburgh.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, 301–305.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013. An SMT Approach to Automatic Annotation of Historical Text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series 18; Linkping Electronic Conference Proceedings*, 54–69.

Rannóg an Aistriúcháin [The Translation Section]. 1945. *Litriú na Gaeilge: Lámhleabhar an Chaighdeáin Oifigiúil [Irish Spelling: Handbook of the Official Standard]*. Oifig an tSoláthair, Baile Átha Cliath.

Rannóg an Aistriúcháin [The Translation Section]. 1958. *Gramadach na Gaeilge agus Litriú na Gaeilge [Grammar of Irish and Spelling of Irish]*. Oifig an tSoláthair, Baile Átha Cliath.

Kevin P. Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the Workshop "Strategies for developing machine translation for minority languages" at LREC 2006*, 103–107.

Kevin P. Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Proceedings of the 3rd Web as Corpus Workshop (WAC3) in Louvain-la-Neuve, Belgium*, 5–15.

Kevin P. Scannell. 2013. Mapping the Celtic Twittersphere. http://indigenoustweets.blogspot.ie/2013/12/mapping-celtic-twittersphere.html. Retrieved May 2, 2014.

Scottish Qualifications Authority. 2009. *Gaelic Orthographic Conventions*. http://www.sqa.org.uk/sqa/files_ccc/SQA-Gaelic_Orthographic_Conventions-G-e.pdf Ughdarras Theisteanas na h-Alba, Glasgow.

Francis M. Tyers. 2009. Rule-based augmentation of training data in BretonFrench statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation, EAMT09*.

Machine translation, the process of automatically translating text from a source lan-guage (e.g. English) to a target language (e.g. French), has achieved impressive results in recent years. However, modern machine translation methods rely heavily on paral-lel data â€“ millions of sentences translated from the source to the target language. Such parallel data is not readily available for most pairs of source and target languages. The goal of this thesis is to explore ways of using other types of data to improve the trans-lations generated by machine translation systems. We consider two main type... In [1], we have proposed systems for text normalization based on statistical machine translation (SMT) methods which are constructed with the support of Internet users and evaluated those with French texts. Internet users normalize text displayed in a web interface in an annotation process, thereby providing a parallel corpus of normalized and non-normalized text. With this corpus, SMT models are generated to translate non-normalized into normalized text. In this paper, we analyze their efficiency for other languages. Additionally, we embedded the English annotation process for training data in Amazon Mechanical Turk and compare the quality of texts thoroughly annotated in our lab to those annotated by the Turkers. Statistical Machine Translation (SMT), as introduced by Brown et al. [5], takes the view that every sentence S in a source language has a possible translation T in the target language. Building on top of this fundamental assumption, SMT based approaches assign to each (S, T) sentence pair the probability P (T |S), which is interpreted as the probability that sentence T is the translated equivalent in the target language of the sentence S in the source language. Accordingly, statistical approaches dene the problem of Machine Translation as: T = arg max P (T |S). T.Â Phrase-based translation models improved the translation quality over IBM models and many re-searches tried to advance the state-of-the-art with these models. We present a statistical model for translation from Scottish Gaelic to Irish that we hope will facilitate communication between the two language communities, especially in social media. An important aspect of this work is to overcome the orthographical differences between the languages, many of which were introduced in a major spelling reform of Irish in the 1940â€™s and 1950â€™s.Â As an additional task, the author casts the text normalisation problem as an SMT problem and applies the statistical models for normalisation of historical Irish text. A large amount of quality texts written by native speakers before standardization are unusable for tasks such as language modelling unless they are converted to the standard form. Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation. The first ideas of statistical machine translation were introduced by Warren Weaver in 1949, including the ideas of applying Claude Shannon's...