Rule-based augmentation of training data in Breton–French statistical machine translation

Francis M. Tyers

Dept. de Llenguatges i Sist. Informàtics,
Universitat d'Alacant
E-03071 Alacant (Spain)

Prompsit Language Engineering
Av. St. Francesc d'Assís, 74, 1r-L
E-03195 l'Altet (Spain)

ftyers@prompsit.com

Abstract

This article describes an initial statistical machine translation system between Breton, a Celtic language spoken in France, and French. It also describes a method for leveraging existing resources from an incomplete rule-based machine translation system to improve the coverage and translation quality of the statistical system by generating expanded bilingual vocabulary lists. Results are presented which show that the use of this method improves the results of the system with respect to both the baseline, and the baseline with a lemma-to-lemma bilingual lexicon.

1 Introduction

Breton is a Celtic language of the Brythonic branch largely spoken in Brittany in the north-west of France. Historically it was spoken only in the northern part of Brittany, *Breizh-Izel* (Lower Brittany). This contrasts with *Breizh-Uhel* (Higher Brittany) which is traditionally Romance-speaking.

Although some sources put the number of native speakers at between 500,000 and 600,000 (Gordon, 2005), a more up-to-date estimate can be found from the organisation *Ya d'ar brezhoneg* which gives a number of 201,083 as of the 11th November 2008, and states that the number is decreasing at a rate of at least one per hour.[1] Breton is classed as a language in serious danger of extinction by the *UNESCO Red Book on Endangered Languages* (Salminen, 1999), a situation exacerbated by the *laissez-faire* policies of the French state.

Like other Celtic languages, Breton exhibits the phenomenon of initial consonant mutation. This occurs when the initial consonant of a word changes based on morpho-syntactic context. For example in the word *tad* "father", the initial consonant mutates to a 'z' (*aspirant* mutation) when the word follows the possessive *ma* "my", so *tad* is "father", while "my father" is *ma zad*.

As for many less-resourced language pairs, while there is little aligned bilingual text, bilingual lexicons are more readily available. One solution would be to use these bilingual lexicons within a rule-based system that makes use of the features found in the bilingual lexicon: part-of-speech, gender, number etc. to try to compensate for the lack of data with some level of generalisation. Even if little parallel data is available, it is still worthwhile to compare any attempt at a more linguistically motivated system with a greater generalising power with a straight-forward, state-of-the-art non-linguistic approach, such as phrase-based statistical machine translation (SMT).

2 Resources

2.1 Parallel corpora

For any language pair, parallel corpora are the scarcest of all resources. In the case of a language with a small population of speakers and no official recognition, the hurdle is even greater. In contrast with Welsh, there are no bilingual parliamentary proceedings that may be used. As an official body for the defence of the Breton language, the *Ofis ar Brezhoneg* is a big producer of Breton translations and we were given the opportunity to access their translation memories, which

[1] http://www.yadarbrezhoneg.com/?article245; Accessed: 11th November, 2008

Corpus	Number of aligned segments
training	27,987
tuning	1,000
devtest	1,000
test	1,000

Table 1: Split of parallel corpus

mostly contain short segments (with an average length of approx. 9 words per segment) largely in the domains of tourism and computer localisation. This results in approx. 285,000 Breton words and 282,073 French words distributed in 31,000 lines. After basic space- and punctuation-based tokenisation, the total number of distinct tokens for Breton was 36,435 and for French was 41,932. These were split into training, tuning and two test sets as described in table 1.[2]

3 A rule-based system

A rule-based MT system for Breton–French is currently being developed inside the Apertium project.[3] Apertium (Armentano-Oller et al., 2006) is an open-source platform for creating rule-based machine translation systems. It was initially designed for closely-related languages, but has also been applied to work with more distant language pairs, such as Welsh–English (Tyers and Donnelly, 2009) and Basque–Spanish. The translation engine in the platform follows a largely shallow-transfer approach. Finite-state transducers are used for lexical processing, first-order Hidden Markov Models (HMMs) and optionally, Constraint Grammars (CGs) based on VISLCG3[4] are used for part-of-speech disambiguation, and multi-stage finite-state based chunking is used for structural transfer.

The current status of the Breton–French system is as follows: the system has a morphological analyser for Breton with approximately 11,000 lemmata (approx. 85% coverage on open-domain text), a bilingual dictionary with 10,797 part-of-speech tagged correspondences between Breton and French, and a very small number of transfer rules (e.g. for concordance and re-ordering within noun phrases, verbal conjugation and pronoun insertion) adapted from the Spanish–French language pair. It is not currently considered a production system as the coverage of the transfer rules is very sparse.

For examples of entries from the morphological analyser and bilingual lexicon, please see figures 1 and 2 respectively. In the morphological analyser, there are two kinds of paradigms (`<par>`) referenced, the first for specifying the initial consonant mutations described above, the second for listing all the morphological forms of a given word along with their analyses. For example, in the case of verbs, a single combination of lemma and paradigm generates between 37 surface forms (for unmutating initial consonants) and 193 (for mutating initial consonants).

4 A statistical phrase-based system

A phrase-based statistical model was trained using the training and tuning sets mentioned above. Although other language model software is frequently used in the literature, the IRSTLM (Marcello et al., 2008) implementation was chosen as it was available and open-source. A 3-gram language model was trained using the French side of the parallel data. The rest of the training process followed the instructions for the baseline system for WMT08, the shared task in the ACL 2008 workshop on statistical machine translation (Callison-Burch et al., 2008). Only a few modifications in the tokeniser provided were necessary, to deal with the *c'h* character in Breton. The training and tuning corpora were tokenised and lowercased to try to alleviate the data sparseness. BLEU scores optimised with the MERT algorithm (Och, 2003) on the tuning set and obtained on the test set are displayed in table 2.

5 Extending the parallel corpus

As the corpus used for training was much smaller than usually used in SMT, there was a problem of coverage. This was aggravated by the fact that Breton is an inflected language and as mentioned previously also exhibits the phenomenon of initial consonant mutation. Such a small corpus is unlikely to contain the majority of frequent surface forms, and almost certainly would not contain the less frequent ones.

To try and alleviate the problem of low coverage of the training data, it was decided to make use of the resources available in the nascent rule-based system described above. Two approaches

[2]The data used in the experiment may be downloaded from `http://elx.dlsi.ua.es/~fran/brfr_OAB_corpus.tgz` and used under the terms of the GNU GPL.

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

[4]`http://visl.sdu.dk/constraint_grammar.html`

```
<e lm="labourat">
  <i>labour</i>
  <par n="labour/at__vblex"/>
</e>
<e lm="kadarnaat">
  <par n="initial-k"/>
  <i>adarna</i>
  <par n="labour/at__vblex"/>
</e>
```

Figure 1: Example of morphological analyser entries for two verbs (*labourat* 'to work' and *kadarnaat* 'to confirm'), including inflectional paradigm (`labour/at__vblex`) and mutation paradigm (`initial-k`)

```
<e>
  <p>
    <l>labourat<s n="vblex"/></l>
    <r>travailler<s n="vblex"/></r>
  </p>
</e>
<e>
  <p>
    <l>kadarnaat<s n="vblex"/></l>
    <r>confirmer<s n="vblex"/></r>
  </p>
</e>
```

Figure 2: Example of bilingual lexicon entries for two verbs. The bilingual lexicon specifies correspondences between lemmata and parts of speech.

were taken. The first was to simply add the bilingual transfer lexicon from the system to the end of the training data. This consisted of 10,797 lemmata. The second was to automatically generate appropriate mappings between all of the surface forms of the given lemmata in the dictionaries of this system.

There has been existing research in this area, for example Dugast et al. (2008) generated a parallel corpus from a rule-based system to train a phrase-based system, and Schwenk (2009) uses an inflected dictionary to produce training data for a statistical system, albeit in a well resourced language pair (French–English).

In order to generate the surface-form mappings, an expansion of all possible surface forms was taken, along with analyses in the Breton morphological analyser. These analyses were then passed through the rest of the Apertium pipeline in order to produce all of the possible translations of surface forms in French. This produces a bilingual inflected dictionary (see figure 3). It is worth mentioning that as a result of the transfer rules, entries for verbs, are generated, where appropriate (e.g. finite verb tenses) with the corresponding subject pronoun in French, and Breton tenses which are not found in French are converted into French tenses (e.g. past habitual is converted to imperfect).

```
mignon,ami
mignoned,amis
vignon,ami
vignoned,amis
dale,retarde
dale,il retarde
labouren,je travaillais
...
```

Figure 3: Example of output from the dictionary expansion and translation – *mignon* 'friend', *dale* 'late' and 'He is delaying' and *labouren* 'I worked'

This 'expanded' bilingual dictionary of surface forms was added to the end of the training corpus, and consisted of 116,514 mappings of inflected Breton forms to inflected French forms.

6 Evaluation and error analysis

As time has not yet been found for a manual evaluation, below are presented the BLEU (Papineni et al., 2002) scores for the three statistical models described above, along with a baseline word-for-word translation generated by the unfinished RBMT system. As expected, the number of unknown words decreases when the bilingual lexicon is added to the training data, and even more when the fully-expanded bilingual lexicon is added. The rise in BLEU (keeping in mind that these are short sentences of ten words on average) is probably also due to side effects such as a better word alignment and a better French context available to the language model scoring. When comparing systems 3 and 4, a quick manual review may attribute most changes to plural forms.

See examples in table 3. The first example shows how the plural form of the Breton for *syllabe* (syllable) could be matched thanks to the morphological extension of the lexicon. In example 2, another kind of extension could be used by the decoder. In French, inflected verbs require the presence of subject pronouns, whereas in Breton this is not the case. This may lead to alignment errors

System	Description	BLEU	Phrase pairs	Unknown words in `devtest`
system 1	word-for-word	0.16	n/a	1,191
system 2	baseline phrase-based SMT	0.29	800k	623
system 3	+ uninflected dictionary	0.30	807k	562
system 4	+ inflected dictionary	0.36	843k	531

Table 2: BLEU scores

Example 1	*Benveg troc'hañ dre silabennoù*
ref	outil de césure par syllabe
gloss	hyphenation tool
system 3	outil coupe par silabennoù
system 4	outil de coupure par syllabes
Example 2	*E rankit kevreañ ouzh an holl darzhioù roadennoù*
ref	Vous devez vous connecter à toutes les sources de données
gloss	You should connect to all of the data sources
system 3	Devez connexion de données . les darzhioù
system 4	Vous devez se connecter tous les darzhioù de données
Example 3	*Emirelezhioù Arab Unanet*
ref	émirats arabes unis
gloss	United Arab Emirates
system 3	émirats arabes unies
system 4	émirats arabes unissez

Table 3: Translation examples

especially with sparse data. In this example, the second plural form in present tense of the Breton verb *rankout* (to have to), *rankit* was mapped to its French equivalent with the corresponding pronoun *vous devez*.

It is also worth noting that the error *se connecter* for *vous connecter* could be alleviated with a more robust verb generation. In French the verb is reflexive, and this is marked in the bilingual lexicon, but the appropriate reflexive pronoun is not yet generated by the rule.

In example 3, the translation of *Emirelezhio Arab Unanet* (United Arab Emirates) displays the adjective for "united" with the incorrect gender. System 4 does not perform better, since it insteads outputs the imperative form of the corresponding verb "unite!". It is very likely that in a real parallel corpora the correct translation (as an adjective) would have been more frequent than the one picked up here in decoding from the extended bilingual lexicon.

7 Conclusions and future work

This paper has presented, to my knowledge, the very first results on Breton to French machine translation. While comparing BLEU scores on a rule-based and a statistical system is not meaningful (Callison-Burch et al., 2006; Labaka et al., 2007), it has shown that the work on the linguistic coding of dictionary entries helped improve a statistical model that had to be trained on little data.

One of the avenues for improving the baseline statistical system would be to add a larger language model on the target side. It would probably also be possible to try to learn probabilities for the rule-based created phrase pairs as in Koehn and Knight (2000). Another option would be to try and create "expanded" phrases based on chunks extracted from a bilingual corpus. For example if you have *war toenn an ti*, "sur le toit de la maison" (on the roof of the house), it would be fairly straightforward given the rule-based system to generate all possible morphological combinations, viz. *war toennoù an ti*, "sur les toits de la maison" (on the roofs of the house), *war toennoù an tiez*, "sur les toits des maisons" (on the roofs of the houses), and *war toenn an tiez*, "sur le toit des maisons" (on the roof of the houses) respectively.

It is also worth noting that at present the Breton–French lexicon in Apertium has only one (gener-

ally the most frequent) translation per word. It would be feasible to generate more than one entry per word, and then score these on language models.

The method described here is knowledge-light, requiring only a morphological analyser, bilingual dictionary and some very basic transfer rules (for verb conjugation) and could be applied to other under-resourced language pairs to improve the coverage of a statistical system where little parallel data is available.

Acknowledgements

I am very grateful to the *Ofis ar Brezhoneg* for making available their translation memory, and for their consistent help during the project. I would also like to extend special thanks to: Fulup Jakez, the director, for his work on verifying and expanding the Breton morphological analyser and Breton–French lexicon, the two contributors to this paper who do not wish to be named, and the reviewers for the helpful comments I received.

References

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

Callison-Burch, Chris, Osbourne, Miles and Koehn Philip 2006. "Re-evaluating the role of Bleu in machine translation research" in *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 249–256

Callison-Burch, Chris, Fordyce, Cameron, Koehn, Philipp, Monz, Christof and Schroeder, Josh 2008. "Further Meta-Evaluation of Machine Translation" in *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 70–106

Dugast, Loïc, Senellart, Jean and Koehn, Philipp 2008. "Can we relearn an RBMT system?" in *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 175–178

Gordon, Raymond G., Jr. (ed.) 2005. *Ethnologue: Languages of the World, Fifteenth edition* (Dallas, Tex.: SIL International)

Koehn, Philip and Knight, Kevin 2000. "Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm" in *Proceedings of the Seventeenth National Conference on Artificial Intelligence* pp. 711–715

Koehn, Philip, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej Constantin, Alexandra and Herbst, Evan 2007. "Moses: Open source toolkit for statistical machine translation" in *ACL 2007, demonstration session*.

Labaka, Gorka, Stroppa, Nicholas, Way, Andy and Sarasola. Kepa 2007. "Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation" in *Machine Translation Summit XI*, Copenhagen, Denmark, pp. 297–304

Federico, Marcello, Bertoldi, Nicola and Cettolo, Mauro 2008. "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models" *Proceedings of the Interspeech 2008*, pp. 1618–1621

Och, Franz J. 2003. "Minimum error rate training in statistical machine translation" *41st Annual Meeting of the Association for Computational Linguistics* pp. 160–167

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu, Wei-jing 2002. "BLEU: a method for automatic evaluation of machine translation" in *40th Annual meeting of the Association for Computational Linguistics* pp. 311–318

Salminen, Tapani 1999. *Unesco Red Book on Endangered Languages*

Schwenk, Holger 2009. "On the use of comparable corpora to improve SMT performance" to appear *EACL-2009*

Tyers, Francis M. and Donnelly, Kevin 2009. "apertium-cy: a collaboratively-developed free RBMT system for Welsh to English" *Prague Bulletin of Mathematical Linguistics* No. 91, pp. 57–66.