Sheffield Systems for the Finnish-English WMT Translation Task

Karin Sim Smith **Lucia Specia** **David Steele**
Department of Computer Science
The University of Sheffield
Sheffield, UK
{kmsimsmith1, l.specia, dbsteele1}@sheffield.ac.uk

Abstract

This paper provides an overview of the Sheffield University submission to the WMT15 Translation Task for the Finnish-English language pair. The submitted translations were created from a system built using the CDEC decoder. Finnish is a morphologically rich language with elements such as nouns and verbs carrying a large number of inflectional types. Consequently, our improvements are based on morphology and include preprocessing steps to handle of morphological inflections inherent in the language, and which otherwise result in lexical sparsity and loss of information.

1 Introduction

This paper outlines The University of Sheffield's submission for the shared translation task, which is part of the 2015 Workshop on Machine Translation. We participated in the Finnish-English language pair task which used news-test-2015 data. 23 systems from 12 organisations took part in this task.

Finnish is an inflectional language containing a productive morphology. The morphological phenomena can lead to a great many inflectional forms. This complex productive morphology can be a barrier to machine translation, with many forms unseen at training. As such, our work was focussed on handling the morphological variation in Finnish with the aim of extracting and transferring as much information as possible - in terms of nominal forms and declensions.

For this paper we describe our baseline system in Section 3, followed by our improvements in Section 4 and potential gains in Section 5. We report our results in Section 6.

2 Related work

In terms of previous work in the translation of morphologically rich languages in MT, Finnish-English has previously featured as a language pair, in the 2005 shared task (Koehn and Monz, 2005).

Chahuneau et al. (2013) experimented specifically with models into morphologically rich languages, we opted to do from Finnish, as a morphologically rich language, into English. Their approach is, however, more systematic, deploying a morphological grammar.

Another approach, used by Ammar et al. (2013) is that of *synthetic translation options*, supplementing the phrase tables to compensate for the sparseness in translating from/to highly inflected languages.

Luong et al. (2010) also investigate morpheme-level extraction, but integrate this into the decoding process itself, instead of the pre-processing step we have. They also incorporate unsupervised morphological analysis and do not rely on language-specific tools, whereas we used a Finnish parser for our morphological analysis

3 Baseline system

For our decoder we used CDEC (Dyer et al., 2010), which essentially is used for rule extraction and decoding. CDEC uses synchronous context-free grammars (SCFGs) as the model for natural language syntax.

The initial tokenization and lower casing were performed using the 'tokenize-anything' and 'lowercase.pl' scripts respectively. They are both included as part of the CDEC suite of tools (similar to those provided with Moses). Fast-align was used to learn the word alignments.

To train the translation model we used the Europarl data set provided. We additionally investigated some of the newly available DCEP corpus (Hajlaoui et al., 2014). This is a resource contain-

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 172–176,
Lisboa, Portugal, 17-18 September 2015.

ing multilingual output from the European Parliament beyond the plenary sessions of Europarl that has recently been made available for use. It contains the parliamentary reports (from the parliamentary committees of the European Parliament), oral and written questions, and press releases, and alignments can be derived for any language pairs in the language matrix. However, through experimentation we discovered issues with misalignments and determined better alignments when isolating the parliamentary reports from the questions and deriving them separately. We subsequently also isolated the press releases and aligned those as well. Although taken as a whole, the system seemed to actually cope with poor alignments - we subsequently experimented with the entire DCEP corpus (Hajlaoui et al., 2014) and achieved comparable results.

The grammar extraction is made up of SCFGs, which generate strings in two languages. The process ultimately builds an SCFG translation grammar (typically from a word-aligned parallel corpus) and in this case is a HIERO grammar . For the purposes of increased speed, per-sentence grammars (PSGs) were used in the translation. PSGs only contain rules that match a single sentence (filtered from larger grammars) and, despite the fact that rules are created for each individual sentence to be translated, they are quickly loadable.

For our Language Model we examined two different approaches. The basic approach purely used the given Europarl dataset, whilst the enhanced approach incorporated a partial selection of monolingual Newscrawl data (provided) taken from the Gigaword corpus in addition to the Europarl data. During experimentation we found that adding the extra Newscrawl data to the language model significantly improved the BLEU score (+0.5). However, due to time constraints we were unable to test this improvement alongside our stemming experiments (Section 4) so did not obtain a compound score (for stemming coupled with the additional monolingual data), which we believe could have been significantly better.

The final output translation initially only had the first letter in every sentence changed to uppercase. The translation was then converted into the SGML format using the 'wrap-xml.perl' script. Unfortunately, just simply converting the initial letter in each of the sentences led to a comparatively poor BLEU-cased score, which we decided had to be improved (see True-casing in section 4)

4 Improvements

4.1 Morphological stemming

Our main improvement to the system was based on the idea that there is a need to deal with the highly inflectional nature of Finnish, as the source language. The fact that Finnish is a morphologically rich language is problematic for machine translation systems. For example, it has 15 grammatical cases which results in a great many declensions of the nouns. This in turn leads to a great deal of lexical sparsity when estimating parameters for the translation model. Ultimately, there is a high incidence of out of vocabulary words, and valuable linguistic information is lost. While inflected forms may have occurred at training time, this means that the simple base form will not necessarily be resolved at decoding time. Even if base forms occurred in training, the inflectional form at decoding time will generally fail to match. The agglutinative nature of Finnish increases the problem further, as many nouns are compounded.

We therefore parsed our data using the Turku Finnish Dependency Parser (Haverinen et al., 2014) which is now available[1]. This parser works efficiently and we were able to process raw input text. The resulting parsed files allowed us to extract the base form of each inflected noun in addition to the parts of speech, dependencies, and grammatical case information. Of this we used the base form and grammatical case information to replace each inflected noun occurring in the test data with its base form, in addition to a place marker for cases deemed relevant. By this we mean that the nominative form, for example, does not result in inflectional variation, nor does it incorporate additional grammatical information, which would be of relevance in English. Often the inflections in Finnish become prepositions in English, so our hope was to retain this additional grammatical information and rely on the case placemarker being aligned to the relevant preposition in English. We decided not to include declensions where the declined form could be ambiguous, and left those unmodified.

We subsequently trained the alignments with our marked up test data. We used the base form and grammatical case information for each noun

[1] http://turkunlp.github.io/Finnish-dep-parser/

occurring in the training and test data and extracted lists of nouns where appropriate.

4.2 Issues encountered with stemming

Of course there is also inflectional variation within some cases, so a strict one-to-one mapping will not necessarily hold true. We did not substitute the base form for plural inflections, which did however result in losses, and which we would attempt to handle better in any future task. More problematic is the fact that a noun can decline in a similar manner for different cases - for example, the word 'kirjan' can be the base form of 'kirja', which means 'book', inflected in the genitive and accusative case - both have the same inflectional form. We dealt with this by only substituting the forms where there was no ambiguity. We determined that this was why we were not seeing improvements that were as significant as we had hoped. In addition, our attempts to rectify the issue were not sufficiently tuned.

Interestingly, we got good results when we only stemmed the nouns in the training set that also appeared in the dev and test sets (not submitted). This suggests that when we stemmed as many nouns as we could, it appeared to do as much harm as good and effectively cancelled itself out.

4.3 Filtering

We attempted an experiment with filtering, based on research proving that the translation direction of the training data makes a significant difference for both the translation model and the language model (Kurokawa et al., 2009; Lembersky et al., 2013; Lembersky et al., 2012). This research indicates a qualitative improvement with much less data. It would seem logical that training on translated data already incorporates some of the crosslingual transfer which is performed by a human translator, and therefore is valuable to capture.

To this end we constructed a directional corpus, filtering the whole of the Europarl for excerpts which were originally in the Finnish language. We did this by tracking the 'language' attribute in the markup to filter out any contributions which had originally been in Finnish. Once we had filtered these out we matched them with corresponding excerpts in the target language, in our case English. One major issue here was that due to the fact that there are only 26 Finnish members of the European Parliament out of a total number of 750, the amount of data that is in Finnish is relatively small. Our resulting filtered data corpus contained just 81,444 lines or sentences. This seemed to prove insufficient to influence the overall score. Unfortunately the DCEP data (Hajlaoui et al., 2014) has no way of determining what the original language was, and thus we had no additional sources for our filtered data.

4.4 True-casing (for BLEU-cased scores)

Due to an initial low BLEU-cased score it was decided that the true-casing had to be enhanced beyond simply capitalising the first letter of each sentence. In addition, time constraints and limited experience with available casing tools led to the creation of a relatively short script in order to improve the casing for the translated sentences. Two simple methods were implemented:

- Firstly, capitalisation statistics (ignoring first words) were taken from the unmodified Europarl corpus and applied to each individual word in the automated translation. For example, there would be instances in the corpus where 'The' appears with a capital 'T' as part of a name, and if this was applied directly then all occurrences of 'the' in the output translation would then be capitalised. Clearly this would not be acceptable and so a ratio of capitalised 'The' versus lower-case 'the' was recorded and if it was over a set limit then all occurrences of 'the' would be capitalised or else none would be. By itself this still has a number of limitations, but it was surprisingly accurate in this case, improving our original BLEU-cased score by nearly +2.0.

- Secondly, each sentence from our automated translation was then cross referenced with its respective sentence in the unmodified source text. Then, for each capitalised word in the source that also appeared in the output translation the capitalisation was carried over and applied. This was particularly effective for items such as place names and other named entities. This second option further enhanced the BLEU-cased score and bought the disparity levels (between cased and non-cased) largely in line with the other submissions (e.g. roughly -1.0).

It should be noted that this approach has its limitations and in the future it is anticipated that ro-

bust, tried and tested tools such as the Moses Truecaser/Recaser will be used to undertake any required casing tasks.

5 Alternative enhancements

5.1 Compound splitting

We also attempted to address the issue of compound splitting, given that Finnish is agglutinative in nature and so has many compound nouns which compact the grammatical inflections. The parsed files usefully gave us the compound forms of our nouns, however, due to lack of time we could not refine our implementation sufficiently.

5.2 Improved Language Model

The primary experimentation of using an enhanced language model that incorporated some of the Newscrawl data showed promising results. Ideally it would have been useful to spend time experimenting with various language models in order to gauge which aspects either positively or adversely affected the output translation. Clearly, for this task the Newscrawl data was largely in domain, and so the full set could have been an appropriate addition to be used in order to further enhance the language model and ultimately produce a more fluid output.

6 Results

Our primary results are displayed in Table 1.

System	BLEU	Cased	TER
Europarl only	12.9	12.3	0.791
Europarl+Newscrawl	13.4	12.5	0.792
Europarl+Stemming	13.4	12.4	0.792

Table 1: Showing the respective BLEU, BLEU-Cased, and Translation Error Rate scores of the three different systems.

Essentially the improvements over the baseline (Europarl only: 12.9) are fairly significant in both cases. This does appear to suggest that extending the language model and applying stemming (separately in this case) are both pertinent enhancements that can be used to improve the overall output translation. However, the fact that the system with fairly extensive stemming is comparable to a standard Europarl system with a slightly enhanced language model highlights a couple of points:

- Further extending the language model should carry significant gains and produce a smoother final translation.
- Stemming has potential, but our methods were a little too simplistic and some of the issues we encountered appeared to cause damage. This suggests using more robust and complex methods to handle the problems and ambiguity could produce much stronger improvements.
- There is potential to combine an extended language model and stemming information in the same system, which again should produce significant improvements.

7 Conclusions

In this paper we presented our submission, which was produced from a system built using the CDEC decoder. Our improvements included preprocessing to deal with morphological variation in Finnish, as the source language, and an attempt at directional filtering. It appeared that as this was our first submission, we were starting from scratch and had significant time consuming groundwork preparation to perform before any enhancements could me made. Ultimately, a number of improvements were made, but the results were not as strong as initially hoped, and we found that ambiguity and other issues encountered during the stemming introduced a degree of damage, which in turn seemed to put a glass ceiling on our BLEU scores. As such, these problems need to be dealt with in a more concrete and elegant manner.

Finally, using a lightly extended (in domain) language model produced a positive result and so there is scope to explore this avenue further. It is anticipated that experimenting with, and managing the language model could well produce significant gains.

References

Waleed Ammar, Victor Chahuneau, Michael Denkowski, Greg Hanneman, Wang Ling, Austin Matthews, Kenton Murray, Nicola Segall, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. 2013. The CMU machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *Proceedings of the Eighth Workshop on Machine Translation*.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into morphologically

rich languages with synthetic phrases. In *Proc. of EMNLP*.

Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. 2010. Cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 7–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Najeh Hajlaoui, David Kolovratník, Jaakko Väyrynen, Ralf Steinberger, and Dániel Varga. 2014. DCEP -digital corpus of the european parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3164–3171.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missil, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531.

Philipp Koehn and Christof Monz. 2005. Shared task: Statistical machine translation between european languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124. Association for Computational Linguistics.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *In Proceedings of MT-Summit XII*, pages 81–88.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Comput. Linguist.*, 38(4):799–825, December.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2013. Improving statistical machine translation by adapting translation models to translationese. *Comput. Linguist.*, 39(4):999–1023, December.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 148–157, Cambridge, MA, October. Association for Computational Linguistics.