Estonian MT System with Morphology

R&D report 30.04.2015

Tilde Eesti OÜ

This work has been financed by the National Programme for Estonian Language Technology.

The goal of the project is to improve Estonian machine translation (MT) system quality by means of integrating Estonian morphology with a statistical MT (SMT) system. To reach this goal in this project, we built support of morphosyntactic patterns and support of Estonian morphosyntactic tagger into LetsMT infrastructure. Then we built two English-Estonian MT systems as the initial experiments to start using Estonian morphology in a MT. One system is a baseline; another system is just like the baseline yet with morphosyntactic tagger included. Then we evaluated the obtained systems using the following methods:

- Automatic BLEU scores
- Human evaluation
- Human evaluation with iBleu

Finally, we evaluate the results obtained and outline possible next steps to be taken.

Data used in system training

The training corpus includes the publicly available DGT-TM[1] – collection of legislative texts of European Union. Both systems are built using the following releases of DGT-TM: DGT-TM 2007, DGT-TM 2011, DGT-TM 2012, DGT-TM 2013 and DGT-TM 2014, the total of 4 156 272 parallel segments. The Estonian part of DGT-TM releases covers the Estonian monolingual data for these MT systems.

Factors in use

Using morphological factors in SMT is innovative. It has been studied and research gives positive expectations (Skadiņš et al, 2010). Using morphology in MT is crucially important to morphologically rich languages such as Estonian.

The phrase-based approach of SMT allows translating source words differently depending on their context by translating whole phrases, whereas target language model allows matching target phrases at their boundaries. However, most phrases in inflectionally rich languages such as Estonian can be inflected in case, number, tense, mood and other morphosyntactic properties, producing considerable amount of variations.

There are hundreds morphology tags for Estonian. The high inflectional variation of target language increases data sparseness at the boundaries of translated phrases, where a language model over surface forms might be inadequate to estimate the probability of target sentence reliably. We introduced an additional language model over disambiguated morphologic tags in the English-Estonian system. The tags contain morphologic properties generated by a statistical morphosyntactic tagger. The order of the tag LM was set to 7, as the tag data has significantly smaller vocabulary compared to word forms.

[1] https://ec.europa.eu/jrc/en/language-technologies/dgt-acquis

System evaluation

BLEU scores

BLEU scores (Papineni et al, 2002) allow automatic evaluation of the systems. After training the two SMT systems, we obtained the following system quality BLEU scores:

Table 1. BLEU scores obtained from Baseline and MORPH SMT system evaluation

SMT system / Type of BLEU	BASELINE	MORPH
Case sensitive	**49.95**	**45.42**
Case insensitive	52.57	47.97

BLEU score is worse in the case of Morph system – not as it was expected. In so high BLEU rates, it is often other factors, which affect the score (synonyms, different punctuation etc). For that reason in this project we perform human evaluation of the MT systems, too.

Human evaluation

We used a ranking of translated sentences relative to each other for manual evaluation of MT systems. This was the official determinant of translation quality used in the 2011 Workshop on Statistical Machine Translation shared tasks (Callison-Burch et al., 2011), and the methodology is described in detail in Skadiņš et al. (2010, section 3.2).

We used the same test corpus in human evaluation as in automatic evaluation, but before doing human evaluation, we filtered out 25% of the data, which was identical, i.e., both MT systems produced identical translations.

The summary of human evaluation results are presented in Table 2 and in Table 3, with and without ties.

Table 2. Results of all human evaluations for both systems including ties

System	BLEU	Average rank in manual evaluation
EN-ET Baseline	49.95 / 52.57	31.56% ± 5.83%
A tie		43.85% ± 6.23%
EN-ET Morph	45.52 / 47.97	24.59% ± 5.40%

Table 3. Results of all human evaluations for both systems excluding ties

System	BLEU	Average rank in manual evaluation
EN-ET Baseline	49.95 / 52.57	56.20% ± 8.31%
EN-ET Morph	45.52 / 47.97	43.80% ± 8.31%

The results are still insufficiently reliable. The fact that 43.85% ± 6.23% of the responses are ties shows that the systems are very close in quality.

Inter-annotator agreement

The methodology and the tool designed for human evaluation allows us to evaluate inter-annotator agreement. Currently the inter-annotator agreement coefficients (Fleiss, 1971; Randolph, 2005) are shown in Table 4.

Table 4. Evaluation of inter-annotator agreement in human evaluation of SMT systems

Type	Result
Fleiss' kappa interpretation:	0.556 (Moderate agreement)
Free kappa interpretation:	0.565 (Moderate agreement)

iBLEU evaluation

When BLEU scores are high and close, and human evaluation is not reliable, we have yet another way of evaluating the systems. iBLEU (Madnani, 2011) is a method which allows a user to drill down into sentences and perform qualitative examination and comparison in a visual and interactive manner by calculating BLEU scores to each individual sentence. Then the user can compare the sentence of interest to that from another system, and focus on sentences with low BLEU scores or on sentences with big BLEU score differences between the results in the translations.

The iBLEU evaluation is currently carried out by 2 professional editors.

Additional steps taken / development of dependencies

To achieve the result of Morphology integration, system training and more in-depth evaluation in the project, the following additional steps had to be taken:

- Developing Part-of-Speech tagger from the free morphological tool.
- Building another factor model within the MT infrastructure to make use of the morphological tags
- Develop additional evaluation methods (iBleu, inter-annotator agreement in human evaluation).

Future work

This is the first step towards using morphosyntatic tags for Estonian MT using LetsMT platform. As the results are not really as expected, more advanced factor models must be sought, developed and used.

References

[1] Madnani N. iBLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems. Proceedings of the Fifth IEEE International Conference on Semantic Computing. 2011.

[2] Papineni K., Roukos S., Ward T., Zhu W., BLEU: a method for automatic evaluation of machine translation, in *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2002

[3] Skadiņš, R., Goba, K., & Šics, V. (2010). Improving SMT for Baltic Languages with Factored Models. In Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 2192 (pp. 125–132). Riga: IOS Press.

[4] Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 22–64, Edinburgh, Scotland.

[5] Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." Psychological Bulletin, Vol. 76, No. 5 pp. 378–382

[6] Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. Paper presented at the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th, 2005. (ERIC Document Reproduction Service No. ED490661)