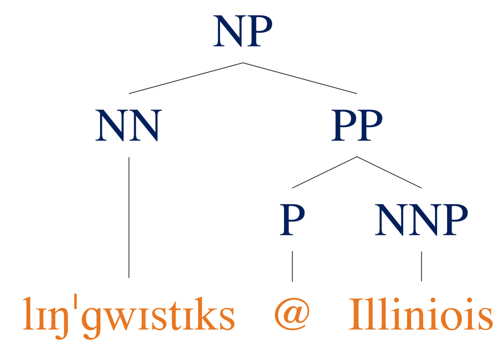
The University of Illinois submission to the WMT 2015 Shared Translation Task



Lane Schwartz, Bill Bryce, Chase Geigle, Sean Massung, Yisi Liu, Haoruo Peng, Vignesh Raja, Subhro Roy, Shyam Upadhyay

University of Illinois at Urbana-Champaign

Abstract

In the 2015 WMT translation task, Finnish-English was introduced as a language pair of competition for the first time. We present experiments examining several variations on a morphologically-aware statistical phrase-based machine translation system for translating Finnish into English. Our system variations attempt to mitigate the issue of rich agglutinative morphology when translating from Finnish into English. Our WMT submission for Finnish-English preprocesses Finnish data with Omorfi (Pirinen, 2015), a Finnish morphological analyzer. We also present results for two other language pairs with morphologically interesting source languages, namely German-English and Czech-English.

1. Methodology

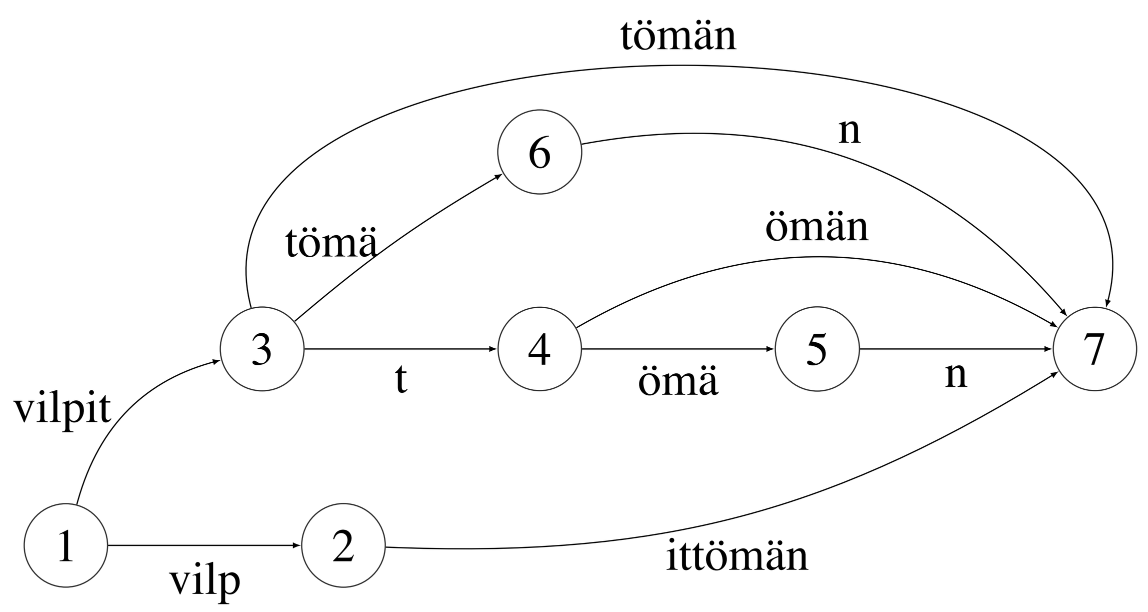
- Use current stable release (v3) of Moses, a state-of-the-art statistical phrase-based machine translation system.
- Train translation models using Europarl (Koehn, 2005), plus Common Crawl corpus and News Commentary (v10) corpus for German-English and Czech-English, and the Wiki Headlines corpus for Finnish-English.
- Train language models on the English Gigaword v5 corpus (Parker et al., 2011) using KenLM (Heafield et al., 2013).

2. Finnish-English

We tried various morphological tokenization schemes on the *source* language (Finnish) in order to mitigate its strong agglutination. The *target* language (English) was tokenized with the default Moses tokenizer script.

2.1 Finnish segmentation using Morfessor

- Adapt lattice technique of Dyer et al. (2009) to Finnish.
- Concatenate source side of training data with its one-best Morfessor (Creutz and Lagus, 2007) segmentation
- Construct source lattices at test time using the top five Morfessor segmentations for each word



An example subgraph in the word lattice that represents the top five segmentations for the Finnish word *vilpittömän*.

Edge weights are calculated according to

$$p(v \mid u, \Theta) = \frac{\sum_{\mathbf{s}:(u,v)\in\mathbf{s}} 2^{\ell_{\mathbf{s}}-\ell_{max}}}{\sum_{\mathbf{s}':(u,v')\in\mathbf{s}'} 2^{\ell_{\mathbf{s}'}-\ell_{max}}},$$

where ℓ_{max} is the highest log likelihood segmentation for the current word. Our Finnish lattice-builder code is available at `https://github.com/smassung/uiuc-wmt15`

Table 1: *Results for Finnish-English using Morfessor*

System	LM	TM	BLEU	-cased
Morfessor	5	8	15.67	14.88
Hiero	6	5	14.99	14.45
Lattice ($n = 2$)	6	8	14.67	14.00
Lattice ($n = 5$)	6	8	14.68	13.95

2.2 Finnish segmentation using Omorfi

First word of Finnish Europarl, as processed by Omorfi:

```
Istuntokauden
Istuntokauden Istunto#kausi N Gen Sg
```

We performed three experimental variations using Omorfi as the morphological segmenter:

1. Segment data using omorfi
2. Concatenate unsegmented and segmented data
3. Segment only out-of-vocabulary words

Table 2: *Results for Finnish-English using Omorfi*

System	LM	TM	BLEU	-cased
Baseline	5	5	16.14	15.25
V1-omorfi	5	5	14.79	14.00
V2-omorfi	5	5	15.14	14.32
V3-omorfi	5	5	**16.90**	**15.98**

3. Czech-English

For Czech-English, we performed experimental variations along two orthogonal dimensions:

- Morphologically segmentation of Czech data
  - no morphological segmentation
  - stem morphemes
  - morphological segmentation using Morfessor
- Part-of-speech (POS) intersection re-ranking feature
  - no POS intersection feature
  - use POS intersection feature to rerank

POS intersection was defined as follows. MorphoDiTa Straková et al. (2014) and the Stanford CoreNLP toolkit Manning et al. (2014) were used to POS tag the Czech and English sentences, respectively. A dictionary maps English and Czech POS tag values. The POS intersection score was defined as the number of identical POS tags between a Czech sentence and the hypothesized English translation.

Table 3: *Results for Czech into English.*

System	BLEU	BLEU-c
Moses trained on Europarl	18.59	17.72
Moses trained on Europarl, Common Crawl and News Commentary	**20.69**	**19.83**
Stemming as pre-processing, Moses trained on Europarl	17.88	17.08
Morfessor trained on Europarl, Moses trained on Europarl	16.48	15.74
POS intersection, Moses trained on Europarl	15.68	13.46
Morfessor trained on Europarl, POS intersection, Moses trained on Europarl	13.43	13.74

4. German-English

Following Holmqvist et al. (2011), we attempt to transform each German sentence de into a sentence de′ with a more English-like word order:

- Parse de using Stanford Parser (Manning et al., 2014)
- Restructure trees using Collins et al. (2005) rules 4 & 6
- Split German compounds using jWordSplitter

Table 4: *Results for German and German′ into English.*

System	BLEU	BLEU-cased	TER
de-en	21.4	22.2	0.938
de′-en	**24.9**	**23.8**	**0.641**

References

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proc. ACL*.

Creutz, M. and Lagus, K. (2007). Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Trans. Speech Lang. Process.*, 4(1):3:1–3:34.

Dyer, C., Setiawan, H., Marton, Y., and Resnik, P. (2009). The University of Maryland Statistical Machine Translation System for the Fourth Workshop on Machine Translation.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proc. ACL*.

Holmqvist, M., Stymne, S., and Ahrenberg, L. (2011). Experiments with word alignment, normalization and clause reordering for SMT between English and German. In *WMT*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proc. MT Summit X*.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proc. ACL*.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English Gigaword Fifth Edition LDC2011T07.

Pirinen, T. A. (2015). Omorfi — Free and open source morphological lexical database for Finnish. In *NODALIDA*.

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proc. ACL*.