SMT Experiments for Commercial Translation of Subtitles

Thierry Etchegoyhen[1], Mark Fishel[2], Jie Jiang[3] & Mirjam Sepesy Maučec[4]

[1] Vicomtech-IK4 [tetchegoyhen@vicomtech.org], [2] TextShuttle [fishel@cl.uzh.ch], [3] CapitaTI [jie.jiang@capita-ti.com], [4] University of Maribor [mirjam.sepesy@uni-mb.si]

Objectives

- Raise the quality of baseline SMT systems built for commercial translation of subtitles, as part of the SUMAT project (www.sumat-project.eu).
- Test state-of-the-art improvement methods & tools on the subtitle domain.
- Perform experiments on the language pairs most likely to improve with the given method.
- Balance: objective MT metrics results - training/decoding efficiency - post-editing effort.

Baselines (2012)

- 14 translation pairs: EN<->DE|ES|FR|NL|PT|SV & SL<->SR.
- Around 800K aligned subtitles per language pair on average (SL-SR exception: 150K) & 13.5 million monolingual subtitles.
- Phrase-based models trained with the Moses toolkit.
- Language models: IRSTLM/KenLM 3|5|6-grams experiments.

Morphosyntactic Factors

- EN <-> ES: no improvement over baseline with any combination of POS & lemmas, with 3|5|7-gram target factor LMs.
- EN -> DE: marginal BLEU increase (+0.2) with POS factors & 7-gram POS LM.
- SL <-> SR: minor improvement (+0.4 BLEU) with POS, lemma & MSD factors. Best scores w/ translation+generation steps.

Named Entities

- EN <-> DE: degraded performance over baseline w/ exclusive and inclusive decoding, complete or reduced NE sets.
- EN <-> SV: no measurable impact in any metric.

Compound Splitting

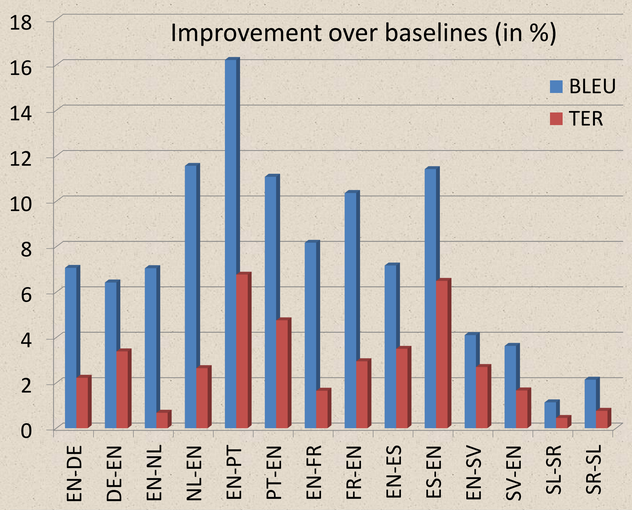
- SV -> EN: No improvement w/ geometric mean of word probabilities segmentation.
- DE -> EN: Statistically significant minor improvement w/ Maximum Entropy on word features (+0.5 BLEU).

Augmented Phrase Tables

- SL <-> SR: No impact of new form FSM generation, form filtering & filled tables w/ form associations.

Parsing

- EN <-> ES: Degraded performance w/ chunk factors on either source or target sides.
- EN <-> DE: Tree-to-tree & String-to-Tree methods showed degraded BLEU performance and equal or slightly worse results on METEOR, TER & Lev5.



Mixture Models

- Additional data:
  - SUMAT: final average of 1.1 million professionally translated aligned subtitles per language pair (exc. SL-SR) and 15.5M total monolingual subtitles.
    - + OpenSubs + Europarl +/- TED +/- EuroparlTV.
- Translation model domain adaptation through perplexity minimization on the SUMAT dev sets + large interpolated/concatenated LMs.
- Several combinations tested: on project's test sets, around 4 BLEU points increase on average with SUMAT corpora included.

	BLEU	TER
EN -> SV	33.0	50.5
SV -> EN	34.3	47.3
EN -> FR	28.1	59.5
FR-> EN	29.4	55.7
EN -> DE	19.7	66.3
DE-> EN	23.2	60.0
EN -> PT	25.8	56.5
PT -> EN	33.1	48.1
EN -> NL	24.3	58.8
NL -> EN	28.0	55.2
SL -> SR	17.8	66.1
SR -> SL	19.1	65.0
EN -> ES	32.5	51.7
ES -> EN	36.0	48.0

Conclusions

- Mixture models as the optimal objective improvement method in terms of metrics and resources, among the approaches that were experimented with.
- Little to no impact with all other tested approaches. Corpora volume increase compensates the minor improvements gained with more complex methods.
- Some error classes (e.g. mistranslated named entities) can induce higher post-editing effort in the subtitling domain. Including more sophisticated analysis steps can thus still be a valid approach in some cases, even in the face of little to no increase of MT metrics.