OpenSubtitles 2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora

Pierre Lison
Norwegian Computing Center (NR)

Jörg Tiedemann
University of Helsinki

Milen Kouylekov
University of Oslo

11th International Conference on Language Resources and Evaluation (LREC 2018)

10/05/2018

Introduction

Movie and TV subtitles are a great resource for compiling parallel corpora:

1. Wide breadth of *linguistic genres*, from colloquial language to narrative and expository discourse.
2. Large databases with millions of subtitles available online, in a wide range of languages
3. Tight coupling between subtitles and their "source material" (a movie or TV episode)

Introduction

- However, the **quality** of the subtitles is often uneven
  - Often created by movie and TV fans
  - Problems with linguistic fluency, faithfulness to the dialogues and adherence to formatting standards
- Sentence alignments from subtitles are also often *less literal* than alignments from other parallel corpora
  - Not direct translations from one another
  - Larger degree of insertions and deletions

Can we automatically estimate *quality scores* for aligned sentence pairs?

Source data

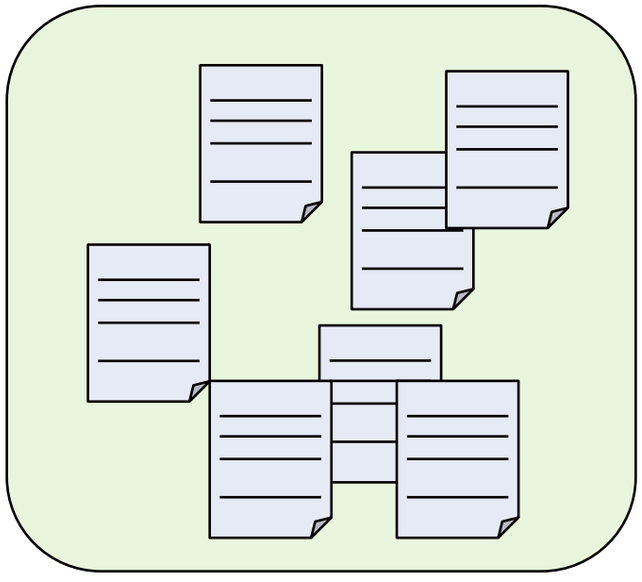
- OpenSubtitles 2018:
  - **3.73 million** subtitles in **60** languages
  - Total of **3.35 billion** sentences (22 billion tokens)
  - Alignment at both document- and sentence-level for all language pairs (1782 bitexts), based on timestamps
- **Preprocessing**:

1. Conversion to Unicode
2. Sentence segmentation
3. Tokenisation
4. OCR error correction
5. Inclusion of meta-data
6. Generation of XML files

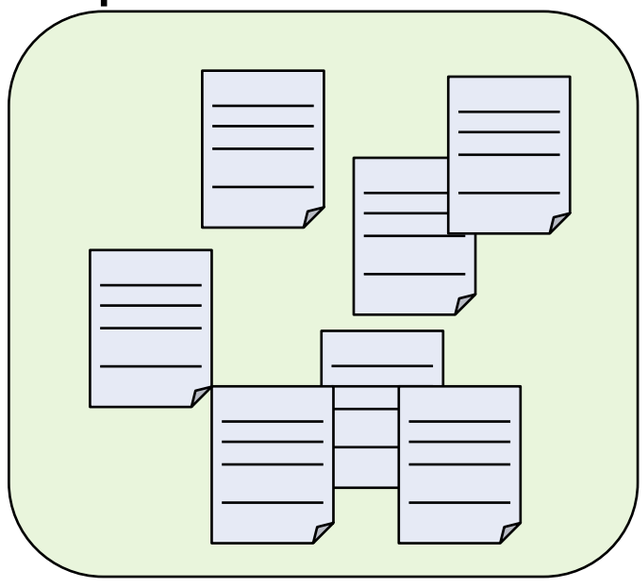
Alignment

- The processed subtitles are then aligned with one another to create a collection of parallel corpora

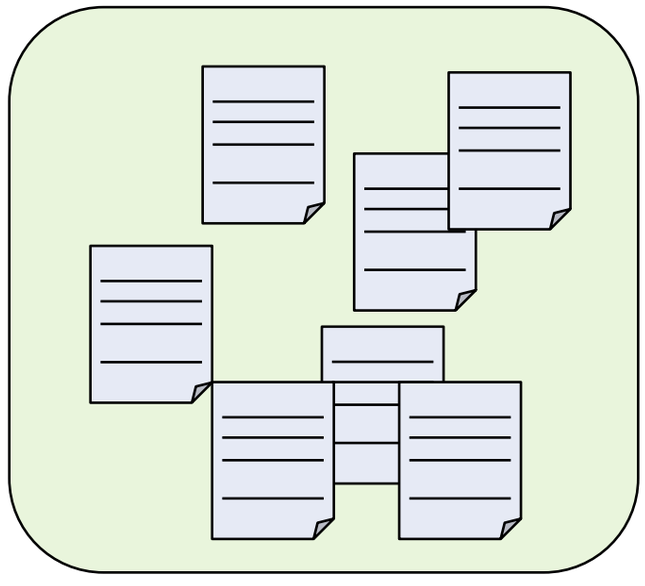
English subtitles



Japanese subtitles



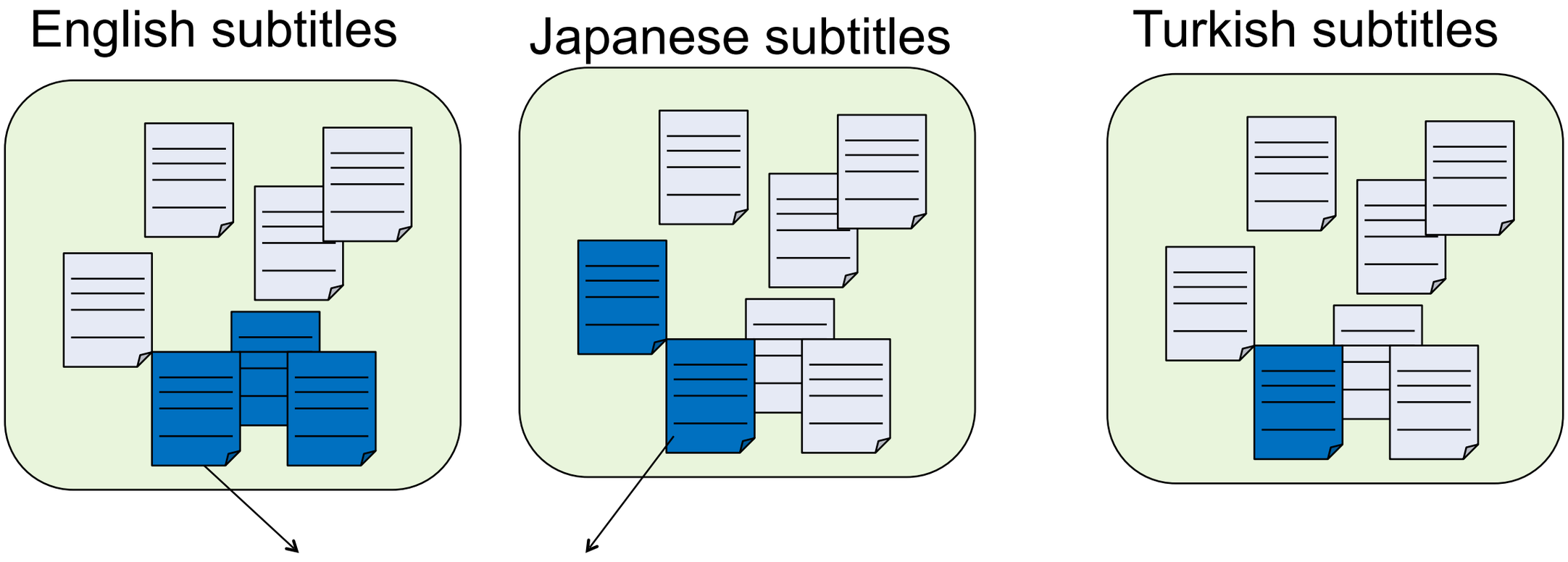
Turkish subtitles



Alignment

= Subtitles for "Love actually" (2003), (using IMDB identifier)

- The processed subtitles are then aligned with one another to create a collection of parallel corpora



Handcrafted scoring function to determine the best subtitle pairs (based on subtitle quality measures + time overlap between the two)

Alignment

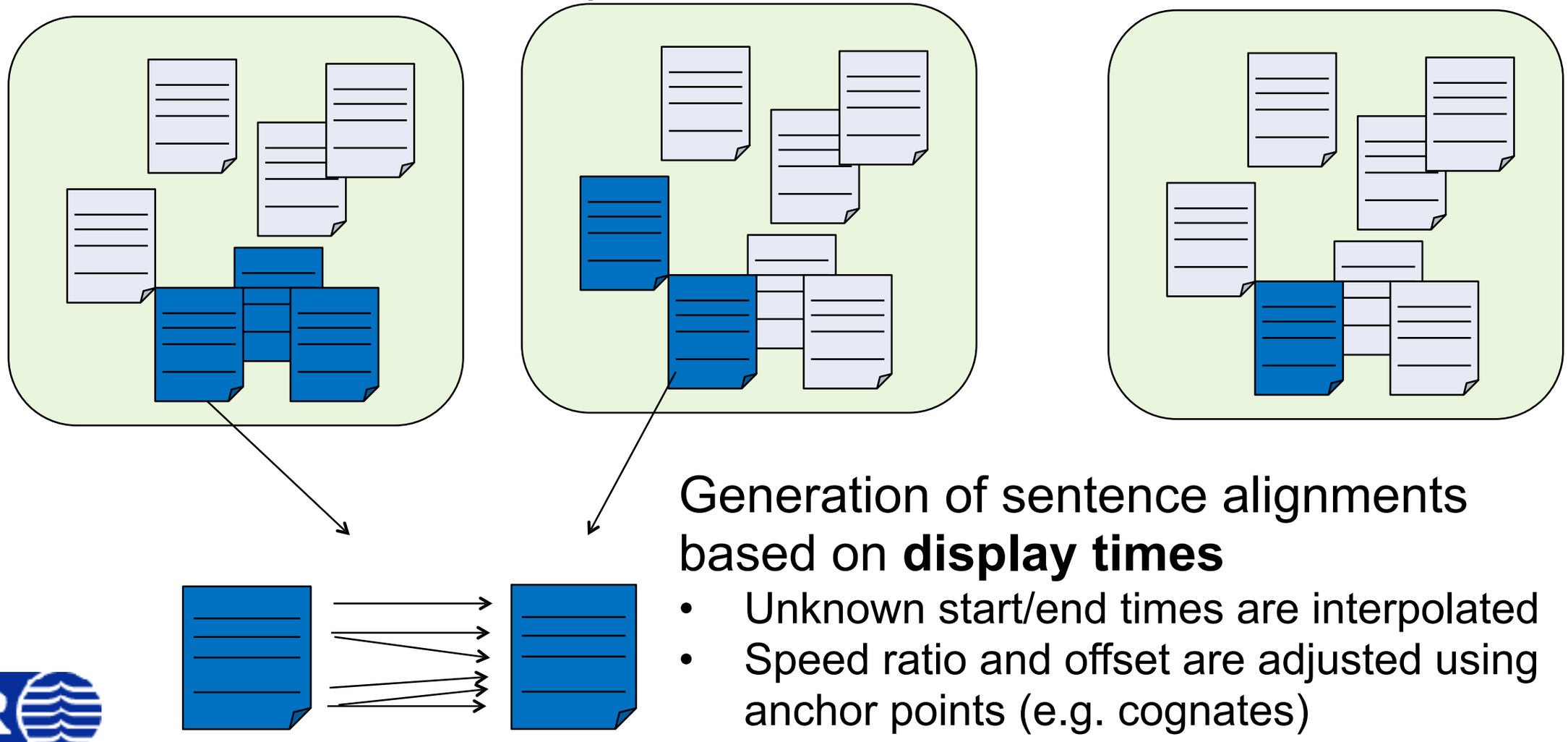
= Subtitles for "Love actually" (2003), (using IMDB identifier)

- The processed subtitles are then aligned with one another to create a collection of parallel corpora

English subtitles

Japanese subtitles

Turkish subtitles



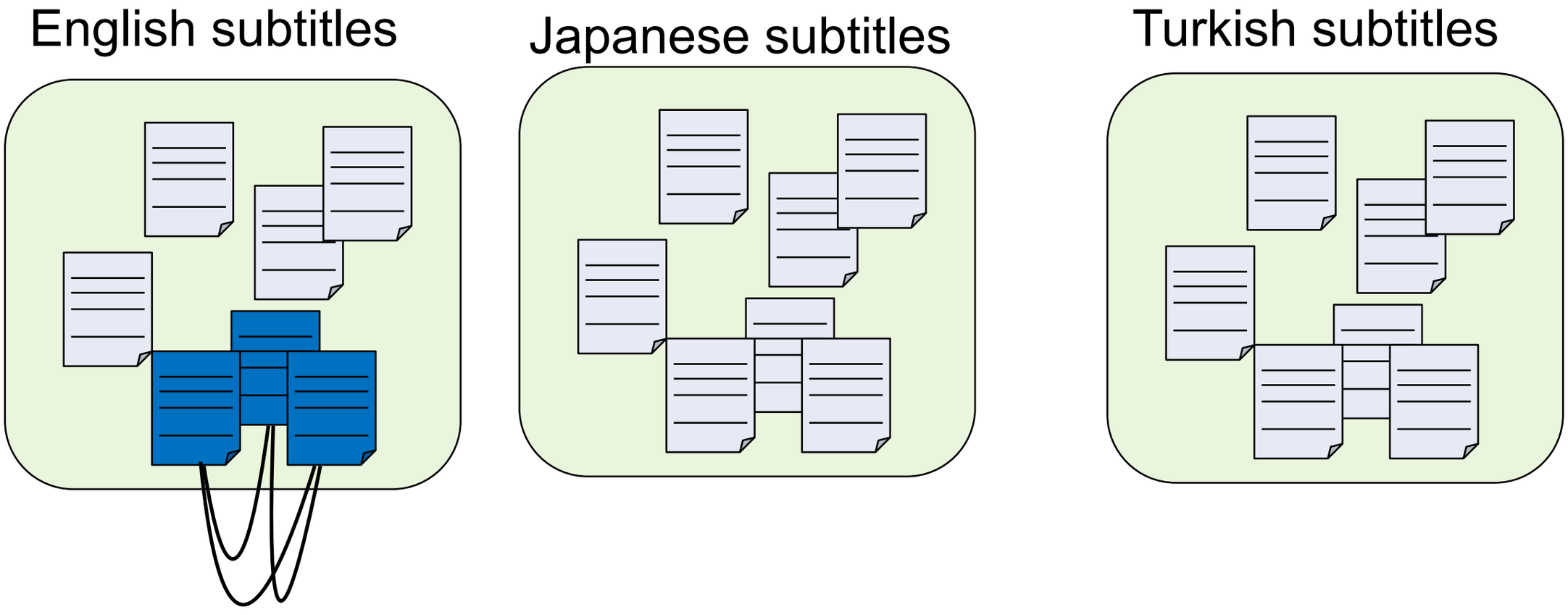
Generation of sentence alignments based on **display times**

- Unknown start/end times are interpolated
- Speed ratio and offset are adjusted using anchor points (e.g. cognates)

Alignment

= Subtitles for
"Love actually" (2003),
(using IMDB identifier)

- The processed subtitles are then aligned with one another to create a collection of parallel corpora



Intra-lingual alignments are also available
(useful to search for e.g. paraphrases)

Scoring model

- **Goal**: learn a regression model q(s_s, s_t) that assigns a numeric quality score to a sentence pair (s_s, s_t)
  - Quality score in the [0,1] range
- **First step**: create a dataset of sentence pairs associated with "gold standard" quality scores
- **Second step**: devise a set of (language-independent) features to be extracted from the sentence pairs
- **Third step**: learn a regression model based on these features and the training set

Measuring alignment quality

- *Key idea*: use IBM Model 1 translation probabilities as a *proxy* for the alignment quality
- Steps:

Compute lexical translation log-probabilities

$$\log P(s|t) = \alpha \sum_{j=1}^{l_s} \log \left(\sum_{i=0}^{l_t} t(s_j|t_i) \right) \qquad \log P(t|s) = \alpha \sum_{j=1}^{l_t} \log \left(\sum_{i=0}^{l_s} t(t_j|s_i) \right)$$

$$\text{score}_{\text{raw}}(s,t) = \min \left(\frac{\log P(t|s)}{l_s}, \frac{\log P(s|t)}{l_t} \right)$$

Normalise for sentence length

$$\text{score}_{\text{final}}(s,t) = \text{scale}_{L_s, L_t} \left(\text{score}_{\text{raw}}(s,t) \right)$$

Rescale per language pair (quantile transform)

Features

Sentence-level features:	Ratio of sentence length (tokens or characters), number of cognates in both source & target, overlap in display times, similar punctuations, etc.
Subtitle-level features:	Number of empty alignments, duration ratio, number of corrected or unknown words, etc.
Meta-level features:	Source & target languages, movie or TV genres, MT translation, user ratings, etc.

- The features are rescaled for each language pair
- Surface features, w/o dependencies on specific resources or tools

Regression model

- *8.3 million sentence pairs* extracted from the OpenSubtitles corpus, covering 760 distinct language pairs.
  - 0.24 % of the total number of sentences in the corpus.
- **Regression models**:
  - Lasso and ridge regression
  - Gradient boosting trees
  - Feedforward neural networks (1 or 2 hidden layers)
- **Evaluation metrics**: (root) mean-square error, and coefficient of determination R^2

Evaluation results

Model	MSE	RMSE	R^2
Baseline (predict mean)	0.009	0.096	0.0
Lasso regression ($\alpha = 0.01$)	0.008	0.092	0.091
Lasso regression ($\alpha = 0.001$)	0.006	0.081	0.303
Ridge regression ($\alpha = 1$)	0.006	0.077	0.356
Gradient boosting (10 regression trees)	0.007	0.085	0.224
Feedforward NN (one hidden layer, dim=100)	0.005	0.071	0.457
Feedforward NN (two hidden layers, dim=100)	**0.005**	**0.070**	**0.470**

Examples of low-quality alignments

Afrikaans:	Kalmeer	*[Calm down]*
Polish:	Dlatego byłem w Wiedniu.	*[That's why I was in Vienna]*

Bosnian:	Tačno tako	*[Exactly]*
Danish:	Og du er tidligere straffet?	*[And you had previous convictions?]*

Greek:	Θεέ μου	*[Oh my god]*
Portuguese:	Residência Mainwaring.	*[Mainwaring Residence.]*

German	(Mystische Musik)	*[(Mystical music)]*
Turkish	Lordum...	*[My Lord...]*

MT experiments

	2016		2018		filtered	
system	subs	news	subs	news	subs	news
en-cs	28.36	12.02	**28.76**	**12.94**	28.35	12.05
en-fi	23.51	11.00	24.00	11.13	**24.12**	**11.49**
en-de	28.71	14.48	**28.92**	**16.07**	**28.92**	14.71
en-ru	23.21	14.21	**23.74**	**15.94**	23.68	15.25
en-tr	**18.67**	6.46	18.58	**7.36**	18.24	6.81
cs-en	38.14	17.18	38.34	**17.26**	**38.37**	16.90
fi-en	26.58	13.80	26.94	10.77	**27.08**	**15.88**
de-en	33.02	18.88	**33.40**	19.16	33.01	**19.24**
ru-en	30.52	18.40	30.15	17.67	**30.58**	**18.71**
tr-en	**25.84**	10.34	25.64	**10.79**	25.32	10.65

- attentional seq2seq model based on Helsinki NMT
- BLEU scores on 2017 subtitles and test data from WMT 2017.

Conclusion

- New major release of the **OpenSubtitles** corpus of movie and TV subtitles
  - 30% increase compared to previous release
  - 3.4 billion sentences, 22.2 billion tokens in 60 languages
- **Quality scoring model** for aligned sentences
  - Combination of sentence-level and global features
  - Can be used to filter out (or assign a lower weight) to sentence pairs with score below a given threshold

Corpus available on OPUS:
http://opus.nlpl.eu/OpenSubtitles2018.php