Cross-lingual Sentence Compression for Subtitles

Wilker Aziz Sheila de Sousa Lucia Specia

w.aziz@wlv.ac.uk
sheilacastilhoms@gmail.com
l.specia@sheffield.ac.uk

May 29, 2012

Outline

1. Motivation
2. Related Work
3. Method
4. Experimental Setup
5. Results
6. Conclusion and Future Work

Scenario

Relevant commercial application

Scenario

Relevant commercial application

- Increasing demand for generation of audiovisual content

Scenario

Relevant commercial application

- Increasing demand for generation of audiovisual content
  - The EC estimated a turnover of **633M€** in 2008[1]

[1] SUMAT: http://www.sumat-project.eu

Scenario

Relevant commercial application

- Increasing demand for generation of audiovisual content
  - The EC estimated a turnover of **633M€** in 2008[1]

[1] SUMAT: `http://www.sumat-project.eu`

- Availability of resources

Scenario

Relevant commercial application

- Increasing demand for generation of audiovisual content
  - The EC estimated a turnover of **633M€** in 2008[1]

    [1] SUMAT: `http://www.sumat-project.eu`
- Availability of resources
  - **54 languages**, 1K bitexts, 8.3G tokens, **1.2G segments**[2]

    [2] OpenSubtitles: `http://opus.lingfil.uu.se`

Problem

Subtitles have to fit

Problem

Subtitles have to fit

- the **space** available on the screen

Problem

Subtitles have to fit

- the **space** available on the screen
- a **time** slot so that they can be read

Problem

Subtitles have to fit

- the **space** available on the screen
- a **time** slot so that they can be read

Native speakers and/or second language learners

Problem

Subtitles have to fit

- the **space** available on the screen
- a **time** slot so that they can be read

Native speakers and/or second language learners

- Audio transcripts (sometimes)

Problem

Subtitles have to fit

- the **space** available on the screen
- a **time** slot so that they can be read

Native speakers and/or second language learners

- Audio transcripts (sometimes)

Foreign viewers

Problem

Subtitles have to fit

- the **space** available on the screen
- a **time** slot so that they can be read

Native speakers and/or second language learners

- Audio transcripts (sometimes)

Foreign viewers

- Translations

Problem

Subtitles have to fit

- the **space** available on the screen
- a **time** slot so that they can be read

Native speakers and/or second language learners

- Audio transcripts (sometimes)

Foreign viewers

- Translations
  - manual (?)

Problem

Subtitles have to fit

- the **space** available on the screen
- a **time** slot so that they can be read

Native speakers and/or second language learners

- Audio transcripts (sometimes)

Foreign viewers

- Translations
  - manual (?)
  - amateur (OpenSubtitles data)

Evidence

Task: English (en) to Brazilian Portuguese (pt)

[1] **D**exter, **H**ow I met your mother and **T**erranova

[2] Text-based sentence alignment

Evidence

Task: English (en) to Brazilian Portuguese (pt)
8K bisegments from recent episodes of 3 TV series[1]

[1] **D**exter, **H**ow I met your mother and **T**erranova

[2] Text-based sentence alignment

Evidence

Task: English (en) to Brazilian Portuguese (pt)
8K bisegments from recent episodes of 3 TV series[1]
Episodes not in OpenSubtitles (same pre-processing[2])

[1] **D**exter, **H**ow I met your mother and **T**erranova

[2] Text-based sentence alignment

Evidence

Task: English (en) to Brazilian Portuguese (pt)

8K bisegments from recent episodes of 3 TV series[1]

Episodes not in OpenSubtitles (same pre-processing[2])

Malformed subtitles

[1] **D**exter, **H**ow I met your mother and **T**erranova

[2] Text-based sentence alignment

Evidence

Task: English (en) to Brazilian Portuguese (pt)
8K bisegments from recent episodes of 3 TV series[1]
Episodes not in OpenSubtitles (same pre-processing[2])
Malformed subtitles

- 33.5% pt: 10 ± 7 chars over

[1] **D**exter, **H**ow I met your mother and **T**erranova
[2] Text-based sentence alignment

Evidence

Task: English (`en`) to Brazilian Portuguese (`pt`)
8K bisegments from recent episodes of 3 TV series[1]
Episodes not in OpenSubtitles (same pre-processing[2])
Malformed subtitles

- 33.5% `pt`: 10 ± 7 chars over
- 36.28% `en`: 8.85 ± 6.73 chars over

[1] **D**exter, **H**ow I met your mother and **T**erranova
[2] Text-based sentence alignment

Evidence

Task: English (`en`) to Brazilian Portuguese (`pt`)
8K bisegments from recent episodes of 3 TV series[1]
Episodes not in OpenSubtitles (same pre-processing[2])
Malformed subtitles

- 33.5% `pt`: 10 ± 7 chars over
- 36.28% `en`: 8.85 ± 6.73 chars over
- 45.2% `pt` longer than `en`: 5 ± 4.5 chars over

[1] **D**exter, **H**ow I met your mother and **T**erranova
[2] Text-based sentence alignment

Evidence

Task: English (en) to Brazilian Portuguese (pt)

8K bisegments from recent episodes of 3 TV series[1]

Episodes not in OpenSubtitles (same pre-processing[2])

Malformed subtitles

- 33.5% pt: 10 ± 7 chars over
- 36.28% en: 8.85 ± 6.73 chars over
- 45.2% pt longer than en: 5 ± 4.5 chars over

Malformed Google translations

[1] **D**exter, **H**ow I met your mother and **T**erranova

[2] Text-based sentence alignment

Evidence

Task: English (en) to Brazilian Portuguese (pt)
8K bisegments from recent episodes of 3 TV series[1]
Episodes not in OpenSubtitles (same pre-processing[2])
Malformed subtitles

- 33.5% pt: 10 ± 7 chars over
- 36.28% en: 8.85 ± 6.73 chars over
- 45.2% pt longer than en: 5 ± 4.5 chars over

Malformed Google translations

- 42.35% pt: 11.6 ± 8.7 chars over

[1] **D**exter, **H**ow I met your mother and **T**erranova
[2] Text-based sentence alignment

Evidence

Task: English (en) to Brazilian Portuguese (pt)
8K bisegments from recent episodes of 3 TV series[1]
Episodes not in OpenSubtitles (same pre-processing[2])

Malformed subtitles

- 33.5% pt: 10 ± 7 chars over
- 36.28% en: 8.85 ± 6.73 chars over
- 45.2% pt longer than en: 5 ± 4.5 chars over

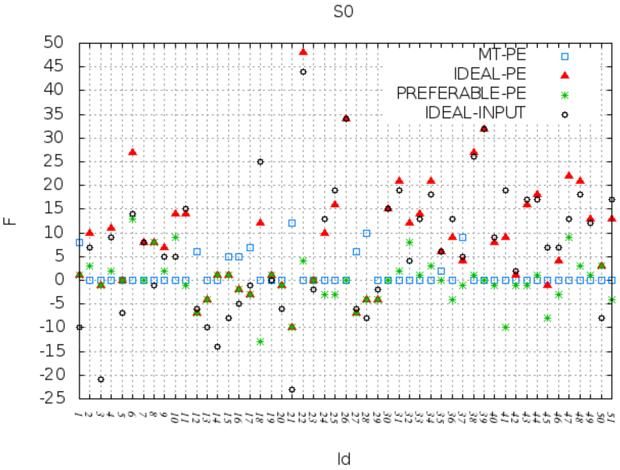
Malformed Google translations

- 42.35% pt: 11.6 ± 8.7 chars over
- 63% pt longer than en: 5.5 ± 4.3 chars over

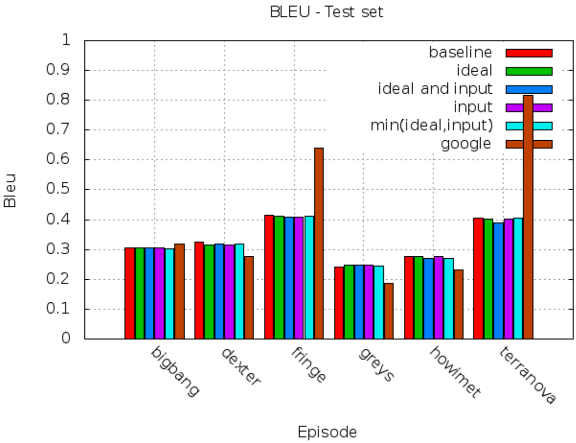
[1] **D**exter, **H**ow I met your mother and **T**erranova
[2] Text-based sentence alignment

Evidence - Reference

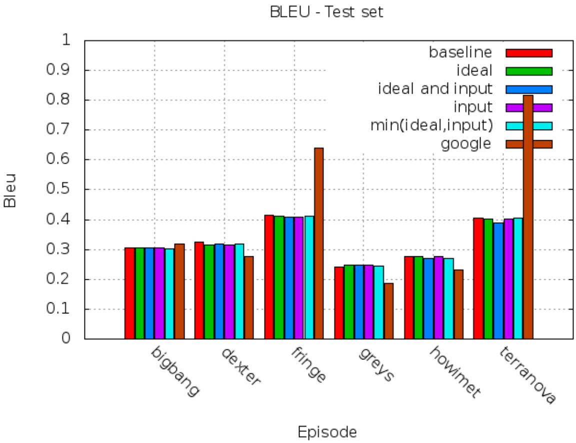
Post-editing/Compression



Evidence - BLEU by show

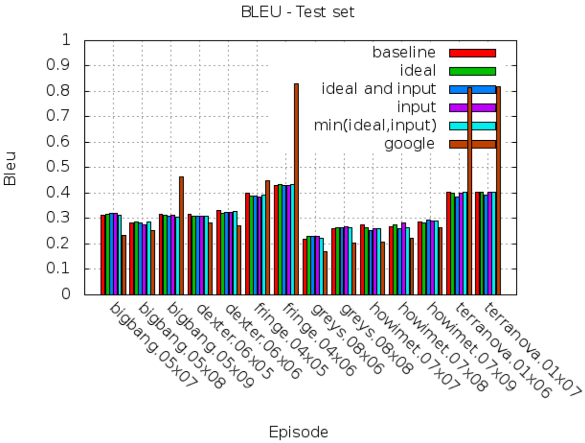


Evidence - BLEU by show



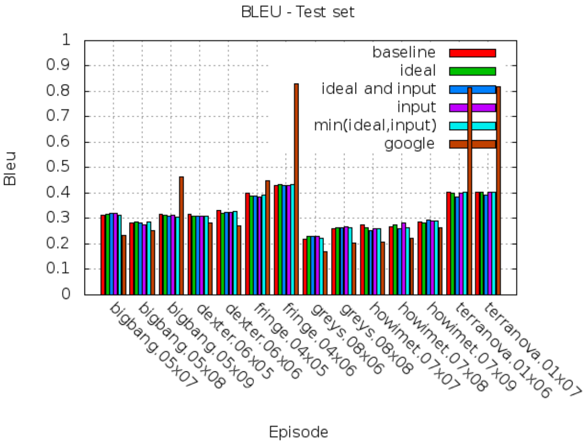
Google: surprising?!

Evidence - BLEU by episode



Wilker Aziz, Sheila de Sousa, Lucia Specia w.aziz@wlv.ac.uk sheilacastilhoms@gmail.com l.specia@sheffield.ac.uk

Evidence - BLEU by episode



Amateur subtitlers do **post-edit**!

Translation for Subtitles

Several attempts

Translation for Subtitles

Several attempts

- Several approaches: RBMT, EBMT and SMT

Translation for Subtitles

Several attempts

- Several approaches: RBMT, EBMT and SMT
- Modest training data

Translation for Subtitles

Several attempts

- Several approaches: RBMT, EBMT and SMT
- Modest training data
- Modest evaluation setups

Translation for Subtitles

Several attempts

- Several approaches: RBMT, EBMT and SMT
- Modest training data
- Modest evaluation setups
  - Small test sets

Translation for Subtitles

Several attempts

- Several approaches: RBMT, EBMT and SMT
- Modest training data
- Modest evaluation setups
  - Small test sets
  - Subjective evaluation

Translation for Subtitles

Several attempts

- Several approaches: RBMT, EBMT and SMT
- Modest training data
- Modest evaluation setups
  - Small test sets
  - Subjective evaluation
  - Few human annotators for evaluation

Compression for Subtitles

Few attempts

Compression for Subtitles

Few attempts

- Handcrafted source-language rules

Compression for Subtitles

Few attempts

- Handcrafted source-language rules
- Word substitution: shorter, but similar (in context)

Compression for Subtitles

Few attempts

- Handcrafted source-language rules
- Word substitution: shorter, but similar (in context)
- Modest evaluation setups

Compression for Subtitles

Few attempts

- Handcrafted source-language rules
- Word substitution: shorter, but similar (in context)
- Modest evaluation setups

Out of the subtitling domain

Compression for Subtitles

Few attempts

- Handcrafted source-language rules
- Word substitution: shorter, but similar (in context)
- Modest evaluation setups

Out of the subtitling domain

- General purpose sentence compression

Compression for Subtitles

Few attempts

- Handcrafted source-language rules
- Word substitution: shorter, but similar (in context)
- Modest evaluation setups

Out of the subtitling domain

- General purpose sentence compression
  "achieve an overall (document-level) compression rate"

Cross-lingual Sentence Compression

Give an SMT system the means to control for length

Cross-lingual Sentence Compression

Give an SMT system the means to control for length

1. Tuning: with adequate dataset, SMT should get closer to reproducing well-formed subtitles

Cross-lingual Sentence Compression

Give an SMT system the means to control for length

1. Tuning: with adequate dataset, SMT should get closer to reproducing well-formed subtitles
2. Model: constrain the length of the output text based on the duration of the subtitle

Cross-lingual Sentence Compression

Give an SMT system the means to control for length

1. Tuning: with adequate dataset, SMT should get closer to reproducing well-formed subtitles
2. Model: constrain the length of the output text based on the duration of the subtitle

Compressing on demand

Length Constraints

Device-dependent norms Cintas and Remael [2007]

Length Constraints

Device-dependent norms Cintas and Remael [2007]

- about 40 chars/line
- maximum of 2 lines/screen
- minimum duration: 1 second
- maximum duration: 6 seconds

Length Constraints

Device-dependent norms Cintas and Remael [2007]

- about 40 chars/line
- maximum of 2 lines/screen
- minimum duration: 1 second
- maximum duration: 6 seconds

Once the device is fixed the constraint is a function of the duration (of the source)

Length Penalty

Phrase-based SMT

Length Penalty

Phrase-based SMT

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k)$$

Length Penalty

Phrase-based SMT

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k)$$

Given a length constraint $c = g(o_f - i_f)$

Length Penalty

Phrase-based SMT

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k)$$

Given a length constraint $c = g(o_f - i_f)$

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) \equiv c - \mathrm{length}(\bar{e}_1^K)$$

Length Penalty

Phrase-based SMT

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k)$$

Given a length constraint $c = g(o_f - i_f)$

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) \equiv c - \mathrm{length}(\bar{e}_1^K)$$

Decomposing

Length Penalty

Phrase-based SMT

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k)$$

Given a length constraint $c = g(o_f - i_f)$

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) \equiv c - \mathrm{length}(\bar{e}_1^K)$$

Decomposing

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k, c)$$

Length Penalty

Phrase-based SMT

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k)$$

Given a length constraint $c = g(o_f - i_f)$

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) \equiv c - \mathrm{length}(\bar{e}_1^K)$$

Decomposing

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k, c)$$

Scaling to a phrase pair

Length Penalty

Phrase-based SMT

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k)$$

Given a length constraint $c = g(o_f - i_f)$

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) \equiv c - \text{length}(\bar{e}_1^K)$$

Decomposing

$$h_{lp}(\bar{f}_1^K, \bar{e}_1^K, c) = \sum_{k=1}^{K} \hat{h}_{lp}(\bar{f}_k, \bar{e}_k, c)$$

Scaling to a phrase pair

$$\hat{h}_{lp}(\bar{f}, \bar{e}, c) = c \times \frac{\text{length}(\bar{f})}{\text{length}(f)} - \text{length}(\bar{e})$$

Informing the Decoder

XML markup

```
<s id="15" lp::ideal="23" lp::input="19" lp::min="19">
  I never felt this .
</s>
```

Data

Data

OpenSubtitles

Data

OpenSubtitles

- Training (unconstrained): 14M segments

Data

OpenSubtitles

- Training (unconstrained): 14M segments

Episodes not in OpenSubtitles (same pre-processing)

Data

OpenSubtitles

- Training (unconstrained): 14M segments

Episodes not in OpenSubtitles (same pre-processing)

- Tuning (unconstrained): 2K segments

Data

OpenSubtitles

- Training (unconstrained): 14M segments

Episodes not in OpenSubtitles (same pre-processing)

- Tuning (unconstrained): 2K segments
- Tuning (constrained): 1,9K (**D**), 1,13K (**H**) and 1,32K (**T**)

Data

OpenSubtitles

- Training (unconstrained): 14M segments

Episodes not in OpenSubtitles (same pre-processing)

- Tuning (unconstrained): 2K segments
- Tuning (constrained): 1,9K (**D**), 1,13K (**H**) and 1,32K (**T**)
- Test (unconstrained): 400 sentences per show

Length Penalties

1. **lp::ideal** is the maximum length, given the duration of the source, so that the subtitle can be read

Length Penalties

1. **lp::ideal** is the maximum length, given the duration of the source, so that the subtitle can be read
2. **lp::input** is the length of the input text

Length Penalties

1. **lp::ideal** is the maximum length, given the duration of the source, so that the subtitle can be read
2. **lp::input** is the length of the input text
   - intuitively one would like to have target subtitles that are close to their source equivalents

Length Penalties

1. **lp::ideal** is the maximum length, given the duration of the source, so that the subtitle can be read
2. **lp::input** is the length of the input text
    - intuitively one would like to have target subtitles that are close to their source equivalents
3. **lp::min** the minimum of the two above

Length Penalties

1. **lp::ideal** is the maximum length, given the duration of the source, so that the subtitle can be read
2. **lp::input** is the length of the input text
   - intuitively one would like to have target subtitles that are close to their source equivalents
3. **lp::min** the minimum of the two above
   - one would like to keep the target length close to the source length while complying with lp:ideal

Systems

Systems

B$_1$: Google

Systems

B$_1$: Google

B$_2$: Moses (unconstrained tuning)

Systems

B$_1$: Google

B$_2$: Moses (unconstrained tuning)

B$_3$: Moses (constrained tuning)

Systems

B$_1$: Google
B$_2$: Moses (unconstrained tuning)
B$_3$: Moses (constrained tuning)
LP$_2$: **B**$_3$ + lp::ideal + lp::input

Systems

B$_1$: Google

B$_2$: Moses (unconstrained tuning)

B$_3$: Moses (constrained tuning)

LP$_2$: **B**$_3$ + lp::ideal + lp::input

LP$_1$: **B**$_3$ + lp::min

Evaluation

Post-editing

8 Brazilian annotators

- PET[3]
- Highlight compliance to length constraints
- Compress only if necessary
- Fix quality only if necessary

HTER Snover et al. [2006]

[3] http://pers-www.wlv.ac.uk/~in1676/pet

Results

Statistical significance in relation to B_3[4]

System	D		H		T	
	TER ↓	LENGTH	TER ↓	LENGTH	TER ↓	LENGTH
B_3	30.3	116.0	20.0	108.5	33.8	120.2
B_1	**63.6**	**156.5**	**52.8**	**144.3**	**63.1**	**152.1**
B_2	**35.7**	**127.3**	**31.3**	**126.9**	**44.1**	**135.8**
LP_2	29.5	115.5	**21.0**	109.1	33.4	**119.3**
LP_1	**28.3**	115.8	20.7	**110.0**	34.8	119.8

[4] x, y and z denote results that are significantly better than a baseline ($p < 0.01$, 0.05 and 0.10, respectively). x, y and z denote results that are significantly worse than a baseline ($p < 0.01$, 0.05 and 0.10, respectively).

Average Length Constraints

Some datasets require more compression

Set	lp::input	lp::ideal	lp::min
D	28.82 ± 15.43	36.99 ± 14.40	26.03 ± 12.86
H	28.40 ± 13.81	33.25 ± 13.77	25.97 ± 12.20
T	28.34 ± 15.22	30.14 ± 11.47	**24.61** ± **11.93**

Malformed Subtitles

Some outputs are easier to compress

Malformed	**B$_1$**	**B$_2$**	**B$_3$**	**LP$_2$**	**LP$_1$**
MT	44.15	34.41	27.40	24.57	25.65
PE	8.50	9.08	7.0	**5.65**	**5.65**

Tuning with Multiple References

Tuning^m: 5 length-compliant reference translations produced for the $1,2K$ test sentences

Test: 600 unseen sentences (200 from each show)

Statistical significance in relation to B$_3^m$[5]

System	TER ↓	LENGTH
B$_3^m$	26.8	103.8
B$_3$	27.0	**106.1**
LP$_2^m$	**26.0**	**103.3**
LP$_1^m$	**25.9**	103.6

[5] x, y and z denote results that are significantly better than a baseline ($p < 0.01$, 0.05 and 0.10, respectively). x, y and z denote results that are significantly worse than a baseline ($p < 0.01$, 0.05 and 0.10, respectively).

Remarks

1. Adequately choosing the tuning data is not only sensible, it actually does a big chunk of the job

Remarks

1. Adequately choosing the tuning data is not only sensible, it actually does a big chunk of the job
2. Controlling the string length can further improve the model's compression capabilities

Remarks

1. Adequately choosing the tuning data is not only sensible, it actually does a big chunk of the job
2. Controlling the string length can further improve the model's compression capabilities
3. Even more in the presence of shorter paraphrases

Remarks

1. Adequately choosing the tuning data is not only sensible, it actually does a big chunk of the job
2. Controlling the string length can further improve the model's compression capabilities
3. Even more in the presence of shorter paraphrases
4. LP models select some nice paraphrases, but they also drop words that are usually added back by human translators: articles, prepositions and conjunctions amongst the most frequent cases

Future Work

1. Explicitly add paraphrases

Future Work

1. Explicitly add paraphrases
2. Model word/phrase deletion

Future Work

1. Explicitly add paraphrases
2. Model word/phrase deletion
3. Decouple translation quality and compression in the evaluation

Future Work

1. Explicitly add paraphrases
2. Model word/phrase deletion
3. Decouple translation quality and compression in the evaluation
4. Additional language pairs

Future Work

1. Explicitly add paraphrases
2. Model word/phrase deletion
3. Decouple translation quality and compression in the evaluation
4. Additional language pairs
5. Use a dataset of professional subtitles

Future Work

1. Explicitly add paraphrases
2. Model word/phrase deletion
3. Decouple translation quality and compression in the evaluation
4. Additional language pairs
5. Use a dataset of professional subtitles **- if we can get one ;)**

Thank you!

References I

Jorge Díaz Cintas and Aline Remael. *Audiovisual Translation: Subtitling. Translation Practice Explained.* St Jerome Publishing, 2007.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, 2006.

Length Constraints

Duration (s)	Length	Ratio (char/s)	Duration (s)	Length	Ratio (char/s)
1.0000	17	17.0000	3.6667	65	17.7273
1.1667	20	17.1429	3.8333	68	17.7391
1.3333	23	17.2500	4.0000	70	17.5000
1.5000	26	17.3333	4.1667	73	17.5200
1.6667	28	16.8000	4.3333	76	17.5385
1.8333	30	16.3636	4.5000	76	16.8889
2.0000	35	17.5000	4.6667	77	16.5000
2.1667	37	17.0769	4.8333	77	15.9310
2.3333	39	16.7143	5.0000	78	15.6000
2.5000	43	17.2000	5.1667	78	15.0968
2.6667	45	16.8750	5.3333	78	14.6250
2.8333	49	17.2941	5.5000	78	14.1818
3.0000	53	17.6667	5.6667	78	13.7647
3.1667	55	17.3684	5.8333	78	13.3714
3.3333	57	17.1000	6.0000	78	13.0000
3.5000	62	17.7143	-	-	-