Phrase-level Quality Estimation for Machine Translation

Varvara Logacheva, Lucia Specia

University of Sheffield

December 3, 2015
Quality Estimation – determination of the quality of an automatically translated segment without reference translation.
Quality Estimation – determination of the quality of an automatically translated segment without reference translation.

Word-level quality estimation:

C’est mon chat. This is my dog
Quality Estimation

Quality Estimation – determination of the quality of an automatically translated segment without reference translation.

**Word-level quality estimation:**

C’est mon chat. This is my dog

OK OK OK BAD

Sentence-level quality estimation:

C’est mon chat. My dog likes chocolate.

This is my cat.

OK

Document-level quality estimation:

Il a besoin de toilettage régulier car le poil du Maine Coon est un poil, mi-long, il ne faut pas être allergique! Il a besoin aussi de shampoings pour sublimer les couleurs (comme le blanc ou le noir par exemple).
Quality Estimation

**Quality Estimation** – determination of the quality of an automatically translated segment without reference translation.

*Word-level quality estimation:*

C’est mon chat.  
This is my dog  
OK  OK  OK  BAD

*Sentence-level quality estimation:*

C’est mon chat.  
My dog likes chocolate.  
This is my cat.
Quality Estimation

**Quality Estimation** – determination of the quality of an automatically translated segment without reference translation.

*Word-level quality estimation:*

C’est mon chat. This is my dog
OK OK OK BAD

*Sentence-level quality estimation:*

C’est mon chat. My dog likes chocolate. BAD
This is my cat. OK
Quality Estimation – determination of the quality of an automatically translated segment without reference translation.

Word-level quality estimation:

C’est mon chat.  
This is my dog  
OK  OK  OK  BAD

Sentence-level quality estimation:

C’est mon chat.  
My dog likes chocolate.  BAD
This is my cat.  OK

Document-level quality estimation:

Il a besoin de toilettage régulier car le poil du Maine Coon est un poil, mi-long, il ne faut pas être allergique! Il a aussi besoin d’un, shampoing tout les mois, et d’un dégraissant tout les deux mois (ou, avant les expositions). Il existe d’ailleurs des shampoings pour, sublimer les couleurs (comme le blanc ou le noir par exemple).

He needs regular grooming for the coat of Maine Coon is a semi-long hair, it should not be allergic! He also needs a shampoo every month, and a degreasing agent every two months (or before exposure). There are also shampoos to sublimate the colors (like white or black for example).
Quality Estimation

Quality Estimation – determination of the quality of an automatically translated segment without reference translation.

**Word-level quality estimation:**

C’est mon chat. This is my dog

OK  OK  OK  BAD

**Sentence-level quality estimation:**

C’est mon chat. My dog likes chocolate. BAD
This is my cat. OK

**Document-level quality estimation:**

Il a besoin de toilletage régulier car le poil du Maine Coon est un poil, mi-long, il ne faut pas être allergique! He needs regular grooming for the coat of Maine Coon is a semi-long hair, it should not be allergic!
Il a aussi besoin d’un, shampoing tout les mois, et d’un dégraissant tout les deux mois (ou, avant les expositions). He also needs a shampoo every month, and a degreasing agent every two months (or before exposure). There are also shampoos to sublime the colors (like white or black for example).
Phrase-level QE: Motivation
Machine Translation errors are not independent

Source: A beautiful flower
Machine Translation: Un bel arbre
Machine Translation errors are not independent

Source: A beautiful flower
Machine Translation: Un bel arbre
Post-edition: Une belle fleur

3 errors
Phrase-level QE: Motivation

Machine Translation errors are not independent

Machine Translation: Un bel arbre
Post-edition: Une belle fleur
Phrase-level QE: Motivation

Machine Translation errors are not independent

Machine Translation: Un bel *arbre*  
Post-editon: Une belle fleur  
Wrong choice of word
Phrase-level QE: Motivation

Machine Translation errors are not independent

Machine Translation:  Un bel arbre
Post-edition:        Une belle fleur

Wrong agreement
Phrase-level QE: Motivation

Machine Translation errors are not independent

Machine Translation: Un bel arbre
Post-editing: Une belle fleur

1 phrase-level error
The majority of Machine Translation systems are phrase-level

Morgen → Tomorrow
fliege → I
ich → will fly
nach Kanada → to the party
zur Konferenz → in Canada
The majority of Machine Translation systems are phrase-level

The translation of "Morgen fliege ich nach Kanada zur Konferenz" is "Tomorrow I will fly to the party in Canada".
The majority of Machine Translation systems are phrase-level

BAD
Phrase-level QE: Challenges

Problems:

- No information on phrase borders
- No definition of “phrase”
- All labels are word-level
- Optimal features for phrase-level QE are unknown
Phrase-level QE: Challenges

Problems:
- No information on phrase borders
- No definition of “phrase”
- All labels are word-level
- Optimal features for phrase-level QE are unknown

Sub-tasks:
- Segment sentences into phrases
- Retrieve phrase-level labels
- Find well-performing phrase-level features
Data segmentation

Available data:

- Source sentence
- Automatically translated sentence
- Post-edition of automatic translation
- Word-level labelling of automatic translation (OK/BAD)

Approach: re-decoding of the data.
Decoder-based segmentation: source-target

Joint segmentation of source and target sentences with constrained decoding
Decoder-based segmentation: source-target

Joint segmentation of source and target sentences with constrained decoding

- Train phrase table on the QE data
Decoder-based segmentation: source-target

Joint segmentation of source and target sentences with **constrained decoding**

- Train phrase table on the QE data
- Decode source side so that the output matches the target side
Decoder-based segmentation: source-target

Joint segmentation of source and target sentences with constrained decoding

- Train phrase table on the QE data
- Decode source side so that the output matches the target side
- Decoder returns phrase segmentation:
Decoder-based segmentation: source-target

Joint segmentation of source and target sentences with constrained decoding

- Train phrase table on the QE data
- Decode source side so that the output matches the target side
- Decoder returns phrase segmentation:

Source: Well ||, things || got even more inappropriate ||!
Target: Bueno ||, las cosas || se pusieron aún más inadecuado ||!
Decoder-based segmentation: source-target

Joint segmentation of source and target sentences with constrained decoding

- Train phrase table on the QE data
- Decode source side so that the output matches the target side
- Decoder returns phrase segmentation:

Source:  Well || , things || got even more inappropriate || !
Target:  Bueno || , las cosas || se pusieron aún más inadecuado || !

Pros
- Correspondence between source and target phrases

Contras
- Phrases are too long
Decoder-based segmentation: target

**Independent decoding** of target side
Decoder-based segmentation: target

**Independent decoding** of target side

- Translate target side with target-to-source SMT system
Decoder-based segmentation: target

**Independent decoding** of target side

- Translate target side with target-to-source SMT system
- Keep phrase segmentation
Decoder-based segmentation: target

**Independent decoding** of target side

- Translate target side with target-to-source SMT system
- Keep phrase segmentation
- Align source and target sides
Decoder-based segmentation: target

**Independent decoding** of target side

- Translate target side with target-to-source SMT system
- Keep phrase segmentation
- Align source and target sides

Well, things got even more inappropriate!

Bueno, las cosas se pusieron aún más inadecuado!
Decoder-based segmentation: target

**Independent decoding** of target side

- Translate target side with target-to-source SMT system
- Keep phrase segmentation
- Align source and target sides

Well, things got even more inappropriate!

Bueno, las cosas se pusieron aún más inadecuado!
Decoder-based segmentation: target

**Independent decoding** of target side

- Translate target side with target-to-source SMT system
- Keep phrase segmentation
- Align source and target sides

Well, || things got || even more || inappropriate ||!

Bueno, || las cosas se pusieron || aún más || inadecuado ||!
Decoder-based segmentation: target

**Independent decoding** of target side

- Translate target side with target-to-source SMT system
- Keep phrase segmentation
- Align source and target sides

Well , || things got || even more || inappropriate || !

Bueno , || las cosas se pusieron || aún más || inadecuado || !

**Pros**
Shorter phrases

**Contras**
Unreliable source phrases:
- depend on alignments
- don’t guarantee source coverage
Labelling

Quality Estimation
Segmentation
Experiments
Conclusions
Source-target
Target
Labelling

- Optimistic: majority labelling
- Mostly good phrase
- Mostly bad phrase
- OK OK BAD OK BAD BAD
- ⇓ ⇓
- OK BAD

- Pessimistic: 30% bad words
- Bad enough phrase
- Not enough bad phrase
- OK OK BAD OK OK OK BAD
- ⇓ ⇓
- BAD OK

- Super-pessimistic: any phrase with a bad word is bad
- Good phrase
- Bad phrase
- Yet another bad phrase
- OK OK BAD OK OK OK BAD OK
- ⇓ ⇓ ⇓
- OK BAD BAD
Labelling

*Optimistic: majority labelling*

[ Mostly good phrase ] [ mostly bad phrase ]

OK OK BAD OK BAD BAD

⇓ ⇓

OK BAD

*Pessimistic: 30% bad words*

[ Bad enough phrase ] [ not enough bad phrase ]

OK OK BAD OK OK OK BAD

⇓ ⇓

BAD OK

*Super-pessimistic: any phrase with a bad word is bad*

[ Good phrase ] [ bad phrase ] [ yet another bad phrase ]

OK OK BAD OK OK OK BAD OK

⇓ ⇓ ⇓

OK BAD BAD
Labelling

**Optimistic: majority labelling**

<table>
<thead>
<tr>
<th>Mostly good phrase</th>
<th>mostly bad phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK</td>
<td>BAD</td>
</tr>
<tr>
<td>OK</td>
<td>BAD</td>
</tr>
<tr>
<td>OK</td>
<td>BAD</td>
</tr>
</tbody>
</table>

**Pessimistic: 30% bad words**

<table>
<thead>
<tr>
<th>Bad enough phrase</th>
<th>not enough bad phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>OK</td>
<td>BAD</td>
</tr>
<tr>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>BAD</td>
<td>OK</td>
</tr>
<tr>
<td>OK</td>
<td>BAD</td>
</tr>
</tbody>
</table>

Varvara Logacheva, Lucia Specia
Labelling

**Optimistic: majority labelling**

- Mostly good phrase: OK, OK, BAD, BAD, BAD, OK
- Mostly bad phrase: OK, BAD, BAD

**Pessimistic: 30% bad words**

- Bad enough phrase: OK, OK, BAD, OK, OK, OK, BAD
- Not enough bad phrase: OK, OK, OK, BAD

**Super-pessimistic: any phrase with a bad word is bad**

- Good phrase: OK, OK
- Bad phrase: OK, BAD, OK
- Yet another bad phrase: OK, OK, BAD, OK
Experimental setup

Language pair: English – Spanish

Datasets:
- WMT-14: 2,000 sentences, manual labelling
- WMT-15: 11,000 sentences, post-edited

Feature sets:
- QuEst sentence-level features
- Word2Vec word embeddings
- QuEst + Word2Vec features

Training methods:
- Conditional Random Fields
- Random Forest
Experimental setup

Language pair: English – Spanish
Experimental setup

Language pair: English – Spanish

Datasets:
- WMT-14: 2,000 sentences, manual labelling
- WMT-15: 11,000 sentences, post-edited
Experimental setup

Language pair: English – Spanish

Datasets:
- WMT-14: 2,000 sentences, manual labelling
- WMT-15: 11,000 sentences, post-edited

Feature sets:
- QuEst sentence-level features
- Word2Vec word embeddings
- QuEst + Word2Vec features
Experimental setup

Language pair: English – Spanish

Datasets:
- WMT-14: 2,000 sentences, manual labelling
- WMT-15: 11,000 sentences, post-edited

Feature sets:
- QuEst sentence-level features
- Word2Vec word embeddings
- QuEst + Word2Vec features

Training methods:
- Conditional Random Fields
- Random Forest
## Segmentation properties

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Target</th>
<th>Source-Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 word</td>
<td><img src="#" alt="Cylinder Chart" /></td>
<td><img src="#" alt="Cylinder Chart" /></td>
</tr>
<tr>
<td>2 words</td>
<td><img src="#" alt="Cylinder Chart" /></td>
<td><img src="#" alt="Cylinder Chart" /></td>
</tr>
<tr>
<td>3 words</td>
<td><img src="#" alt="Cylinder Chart" /></td>
<td><img src="#" alt="Cylinder Chart" /></td>
</tr>
<tr>
<td>4 words</td>
<td><img src="#" alt="Cylinder Chart" /></td>
<td><img src="#" alt="Cylinder Chart" /></td>
</tr>
<tr>
<td>5 words</td>
<td><img src="#" alt="Cylinder Chart" /></td>
<td><img src="#" alt="Cylinder Chart" /></td>
</tr>
</tbody>
</table>

Varvara Logacheva, Lucia Specia
Phrase-level Quality Estimation
13/21
Labelling properties

- Optimistic
- Pessimistic
- Super-pessimistic

Source-target
Target

"BAD" to "OK", %
"OK" to "BAD", %
Optimal parameters: labelling

![Graph showing F1 OK and F1 BAD for different parameter settings.]

- Pessimistic
- Super-pessimistic

Varvara Logacheva, Lucia Specia
Phrase-level Quality Estimation
Optimal parameters: features

F1 OK vs F1 BAD

- QuEst features
- Word2Vec features
- QuEst+Word2Vec features
Optimal parameters: segmentation and training algorithms
Comparison with word-level systems

WMT-14 systems

<table>
<thead>
<tr>
<th>System</th>
<th>$F_1$-BAD</th>
<th>$F_1$-OK</th>
<th>Weighted F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase-wmt-14</td>
<td>62.76</td>
<td>39.07</td>
<td>56.80</td>
</tr>
<tr>
<td>Baseline-all-bad</td>
<td>52.52</td>
<td>0.0</td>
<td>18.7</td>
</tr>
<tr>
<td>FBK-UPV-UEDIN</td>
<td>48.72</td>
<td>69.33</td>
<td>61.99</td>
</tr>
<tr>
<td>LIG</td>
<td>44.47</td>
<td>74.09</td>
<td>63.54</td>
</tr>
</tbody>
</table>
### Comparison with word-level systems

#### WMT-15 systems

<table>
<thead>
<tr>
<th>System</th>
<th>$F_1$-BAD</th>
<th>$F_1$-OK</th>
<th>Weighted F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase-wmt-15</td>
<td>51.84</td>
<td>49.38</td>
<td>51.08</td>
</tr>
<tr>
<td>UAlacant</td>
<td>43.12</td>
<td>78.07</td>
<td>71.47</td>
</tr>
<tr>
<td>SHEF-word2vec</td>
<td>38.43</td>
<td>71.63</td>
<td>65.37</td>
</tr>
<tr>
<td>Baseline-all-bad</td>
<td>31.75</td>
<td>0.0</td>
<td>5.99</td>
</tr>
<tr>
<td>Baseline</td>
<td>16.78</td>
<td>88.93</td>
<td>75.31</td>
</tr>
</tbody>
</table>
Conclusions

- phrase-level QE model for MT
Conclusions

- phrase-level QE model for MT
- decoder-based sentence segmentation techniques
Conclusions

- phrase-level QE model for MT
- decoder-based sentence segmentation techniques
- features: sentence-level and word embeddings
Conclusions

- phrase-level QE model for MT
- decoder-based sentence segmentation techniques
- features: **sentence-level** and word embeddings
- training methods: **CRF** and Random Forest
Conclusions

- phrase-level QE model for MT
- decoder-based sentence segmentation techniques
- features: **sentence-level** and word embeddings
- training methods: **CRF** and Random Forest
- phrase-level systems outperform word-level systems
Thank you