Phrase-level Quality Estimation for Machine Translation

Varvara Logacheva, Lucia Specia

University of Sheffield

December 3, 2015

Quality Estimation

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

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

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**

Quality Estimation

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

Word-level quality estimation:

	This	is	my	dog
C'est mon chat.	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

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 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).	OK

Phrase-level QE: Motivation

Phrase-level QE: Motivation

Machine Translation errors are not independent

Source:	A beautiful flower
Machine Translation:	Un bel arbre

Phrase-level QE: Motivation

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	*Wrong choice of word*
Post-edition:	Une belle fleur	

Phrase-level QE: Motivation

Machine Translation errors are not independent

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

Phrase-level QE: Motivation

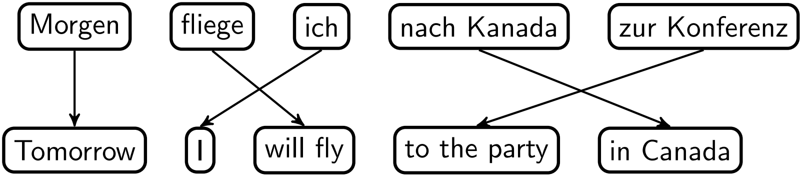
Machine Translation errors are not independent

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

1 phrase-level error

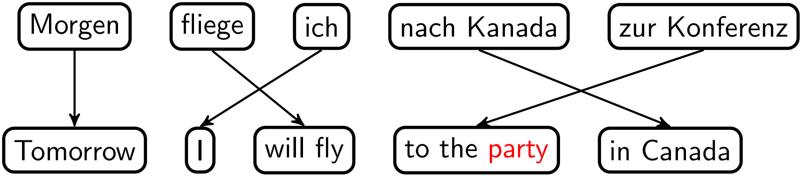
Phrase-level QE: Motivation

The majority of Machine Translation systems are phrase-level



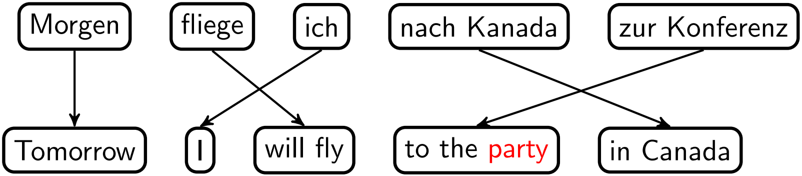
Phrase-level QE: Motivation

The majority of Machine Translation systems are phrase-level



Phrase-level QE: Motivation

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

Labelling

Optimistic: majority labelling

[Mostly	good	phrase]	[mostly	bad	phrase]
OK	OK	BAD	OK	BAD	BAD
	$\Downarrow$			$\Downarrow$	
	OK			**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**		

Labelling

Optimistic: majority labelling

[Mostly	good	phrase]	[mostly	bad	phrase]
OK	OK	BAD	OK	BAD	BAD

$\Downarrow$ **OK** $\Downarrow$ **BAD**

Pessimistic: 30% bad words

[Bad	enough	phrase]	[not	enough	bad	phrase]
OK	OK	BAD	OK	OK	OK	BAD

$\Downarrow$ **BAD** $\Downarrow$ **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

$\Downarrow$ **OK** $\Downarrow$ **BAD** $\Downarrow$ **BAD**

Experimental setup

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

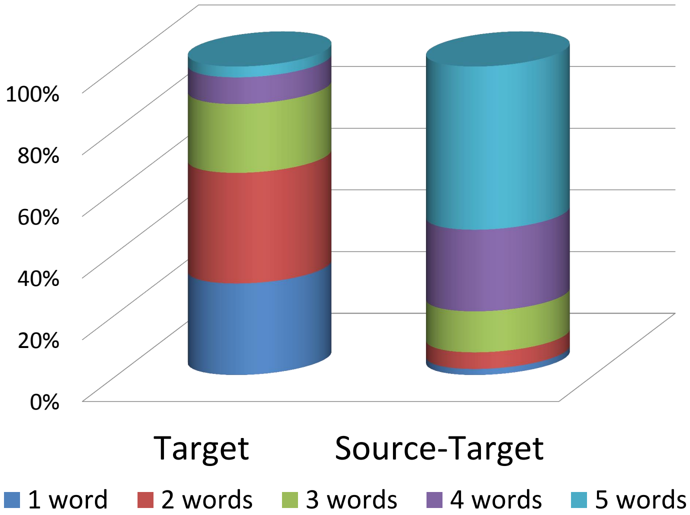
Feature sets:

- QuEst sentence-level features
- Word2Vec word embeddings
- QuEst + Word2Vec features

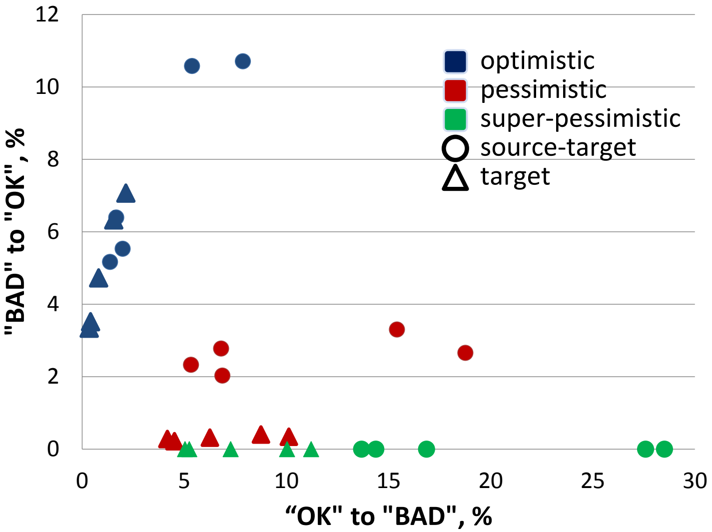
Training methods:

- Conditional Random Fields
- Random Forest

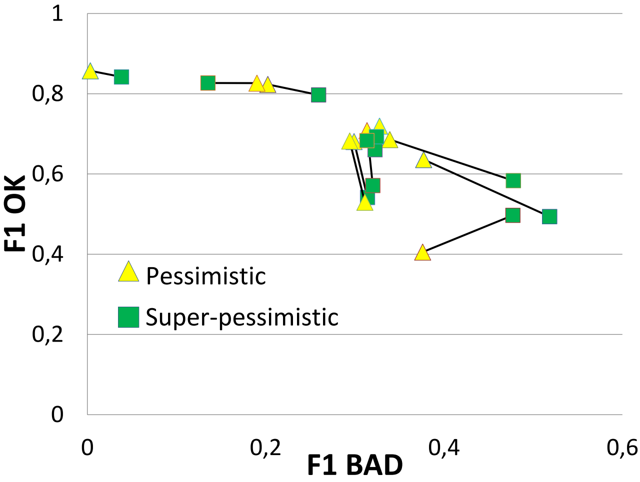
Segmentation properties



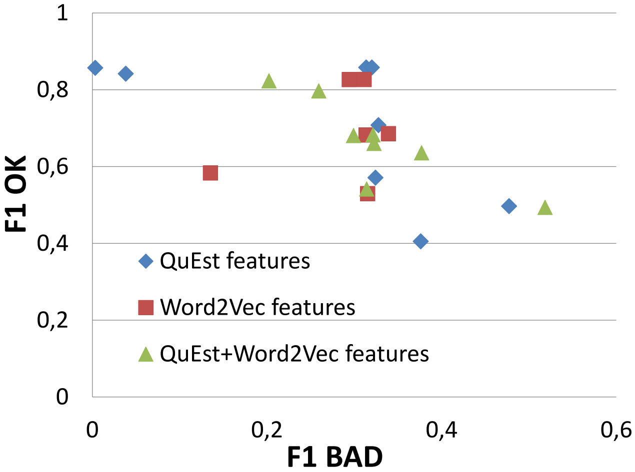
Labelling properties



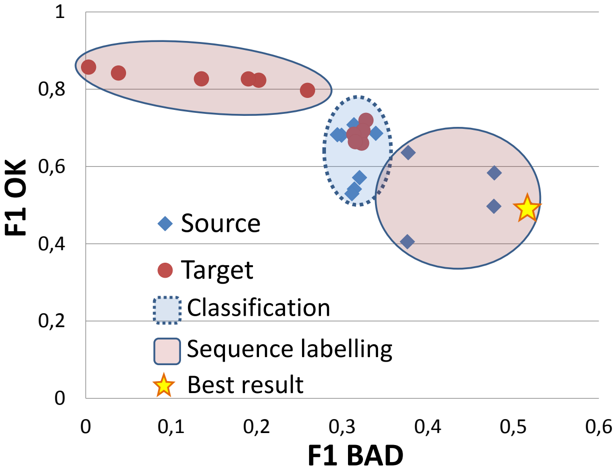
Optimal parameters: labelling



Optimal parameters: features



Optimal parameters: segmentation and training algorithms



Comparison with word-level systems

WMT-14 systems

System	F_1-**BAD** ↑	F_1-OK	Weighted F1
phrase-wmt-14	62.76	39.07	56.80
Baseline-all-bad	52.52	0.0	18.7
FBK-UPV-UEDIN	48.72	69.33	61.99
LIG	44.47	74.09	63.54

Comparison with word-level systems

WMT-15 systems

System	F_1-**BAD** ↑	F_1-OK	Weighted F1
phrase-wmt-15	51.84	49.38	51.08
UAlacant	43.12	78.07	71.47
SHEF-word2vec	38.43	71.63	65.37
Baseline-all-bad	31.75	0.0	5.99
Baseline	16.78	88.93	75.31

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