FONDAZIONE
BRUNO KESSLER

Mining Parallel Fragments from Comparable Texts

Mauro Cettolo, **Marcello Federico**, **Nicola Bertoldi**

Human Language Technologies Research Unit
FBK-irst, Trento - Italy

IWSLT 2010 - Paris, 3 December 2010

Mining Comparable News

Problem

- SMT needs sufficient translated text as training material
- large parallel corpora do not exist for many language pairs
- $\longrightarrow$ try to exploit comparable news, which are easier to collect

Comparable news

news in different languages with the same content,
which are not direct translations of each other

Parallel texts in comparable news

General Scheme

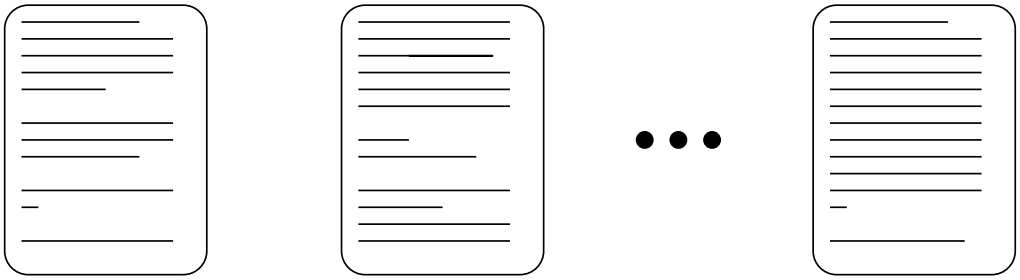
Previous work on how to collect parallel data from comparable data:

1. cluster multilingual documents, by metadata, heuristics, IR ..
2. split documents into sentences
3. pair sentences across documents, by length, lexical overlap, word alignment ..
4. filter sentence or fragment pairs which align very well

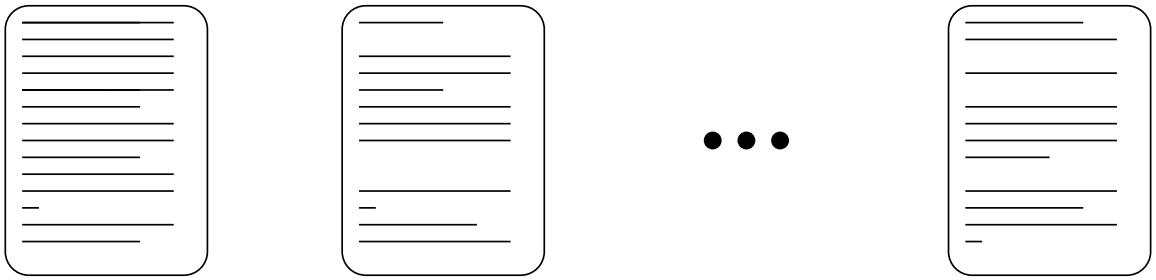
Our approach fits this scheme and it is new on some aspects.

General Scheme

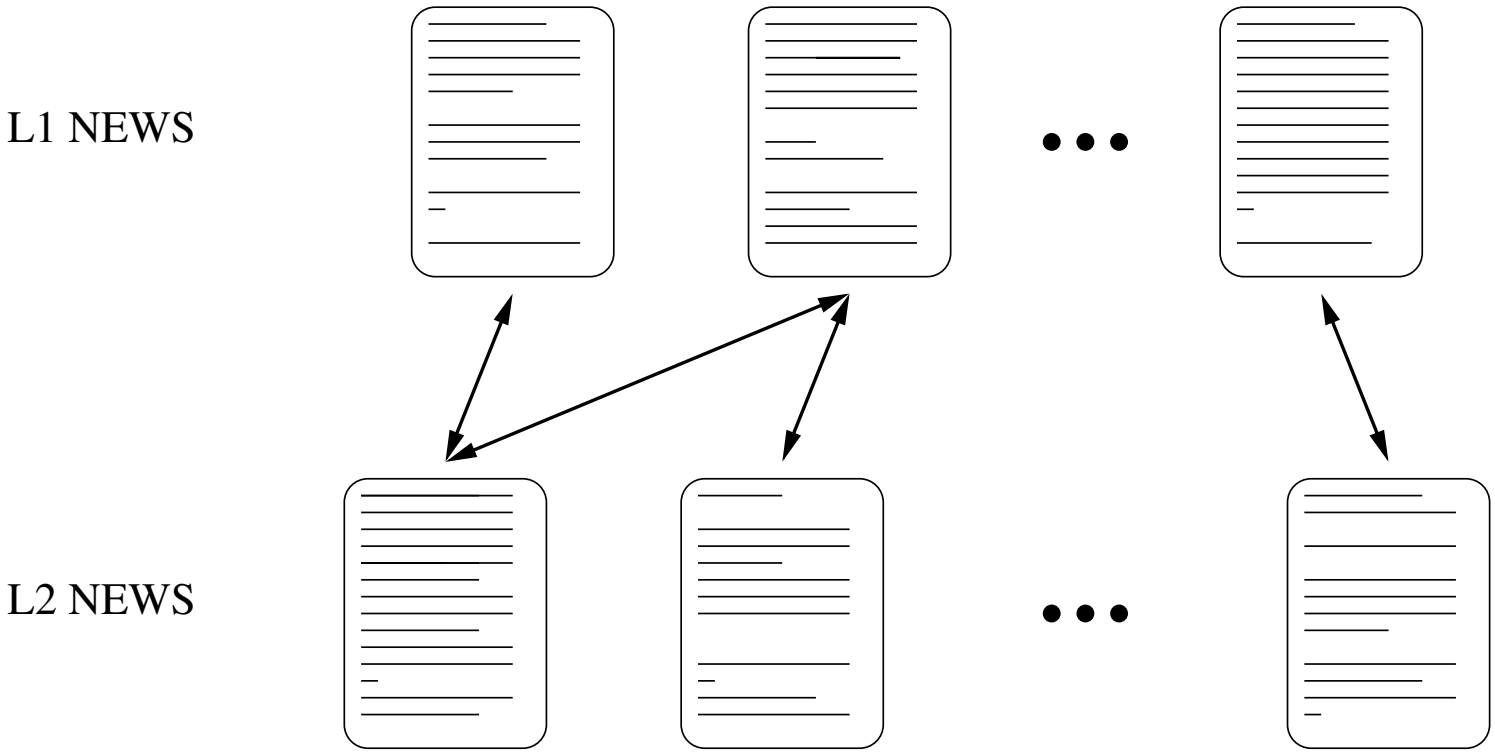
L1 NEWS



L2 NEWS

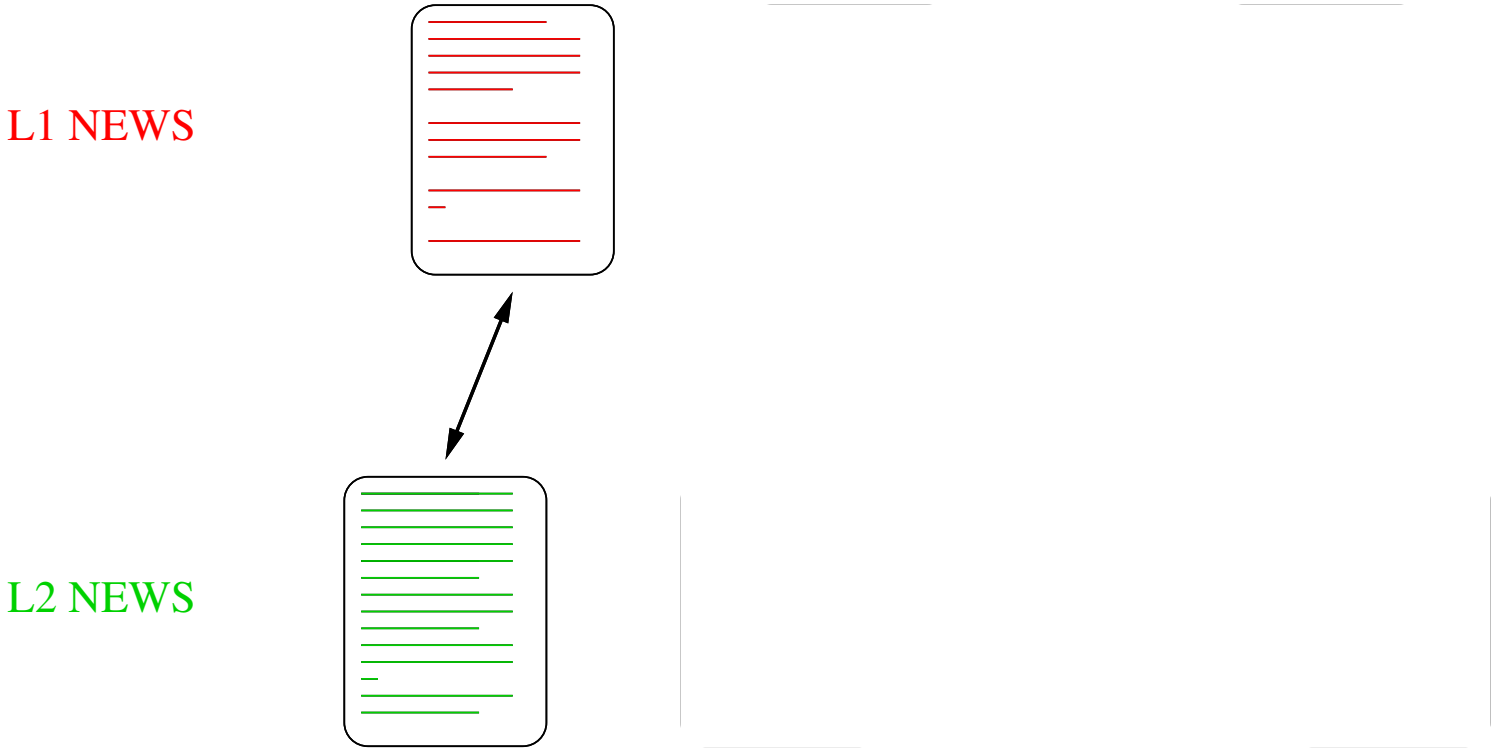


General Scheme: Pairing of Documents



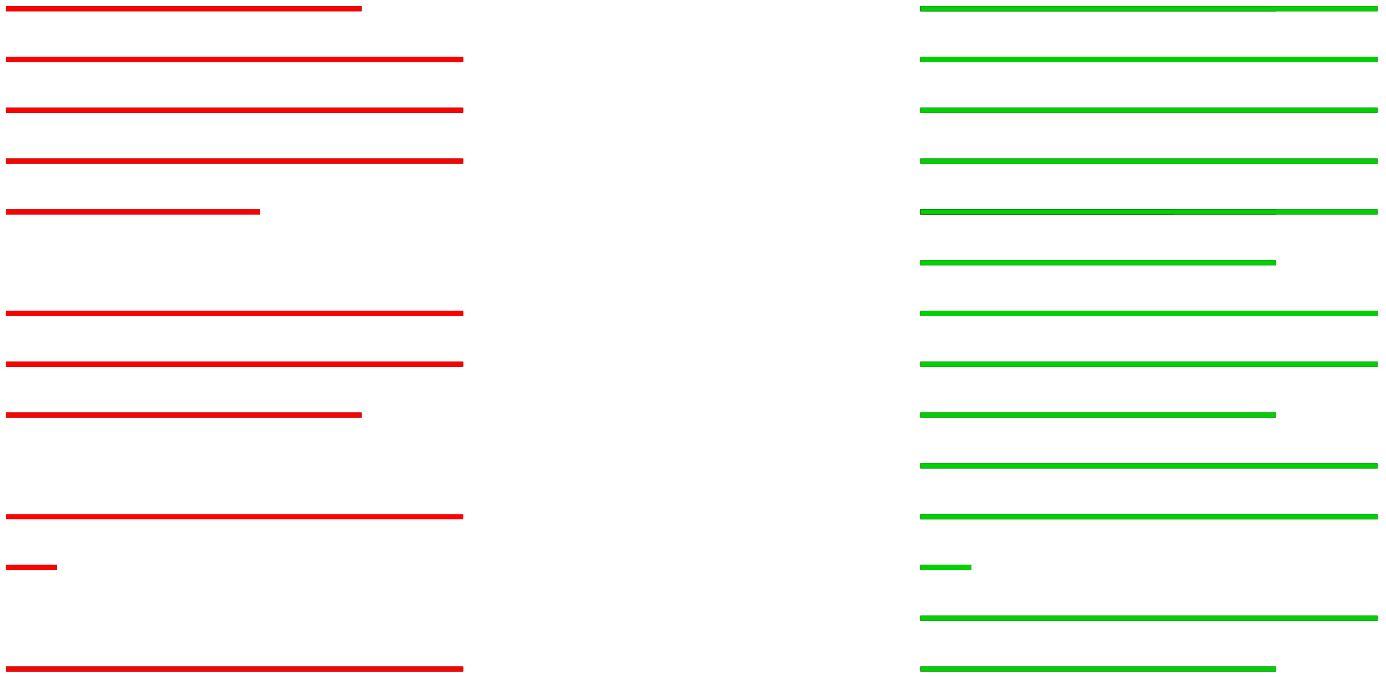
- by exploiting metadata (images, explicit links)

General Scheme: Sentence Splitting

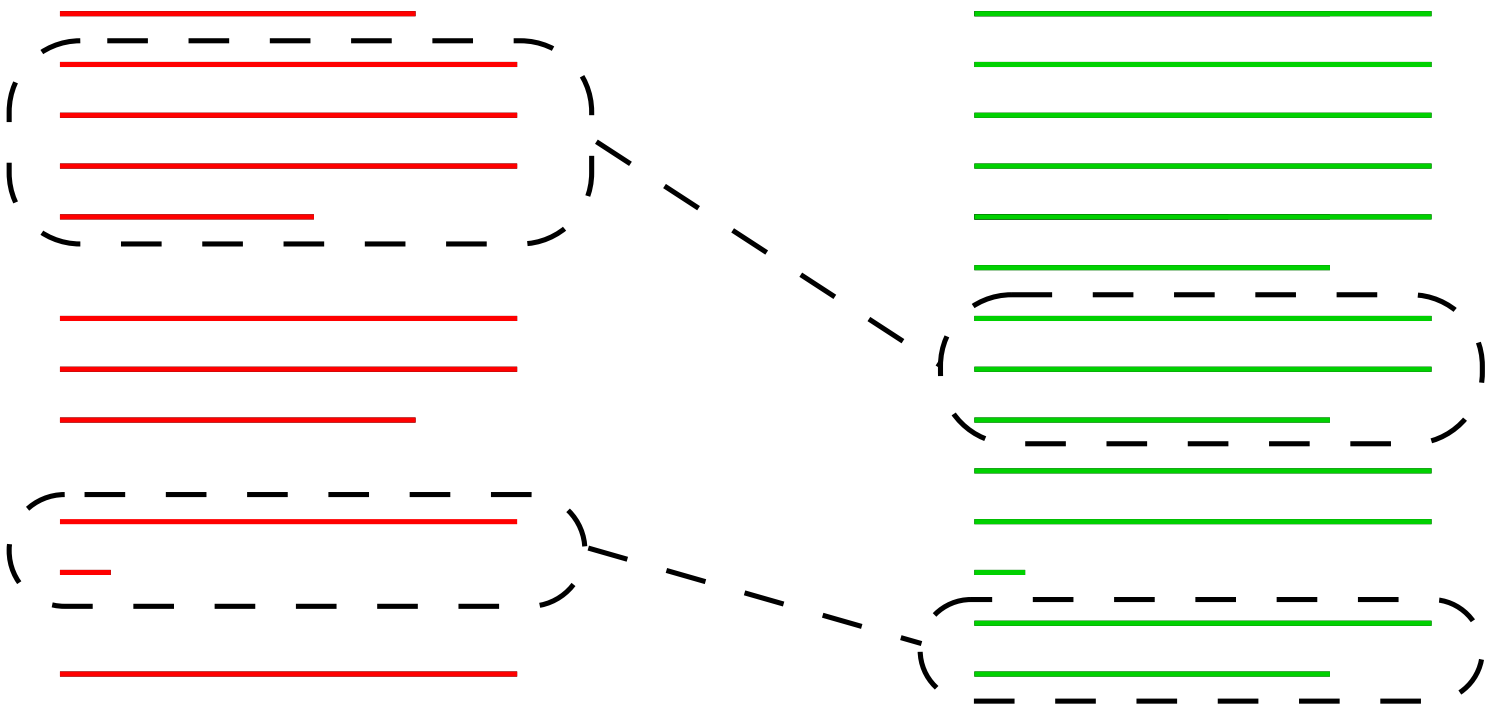


- on strong punctuation

General Scheme: Pairing of Sentences

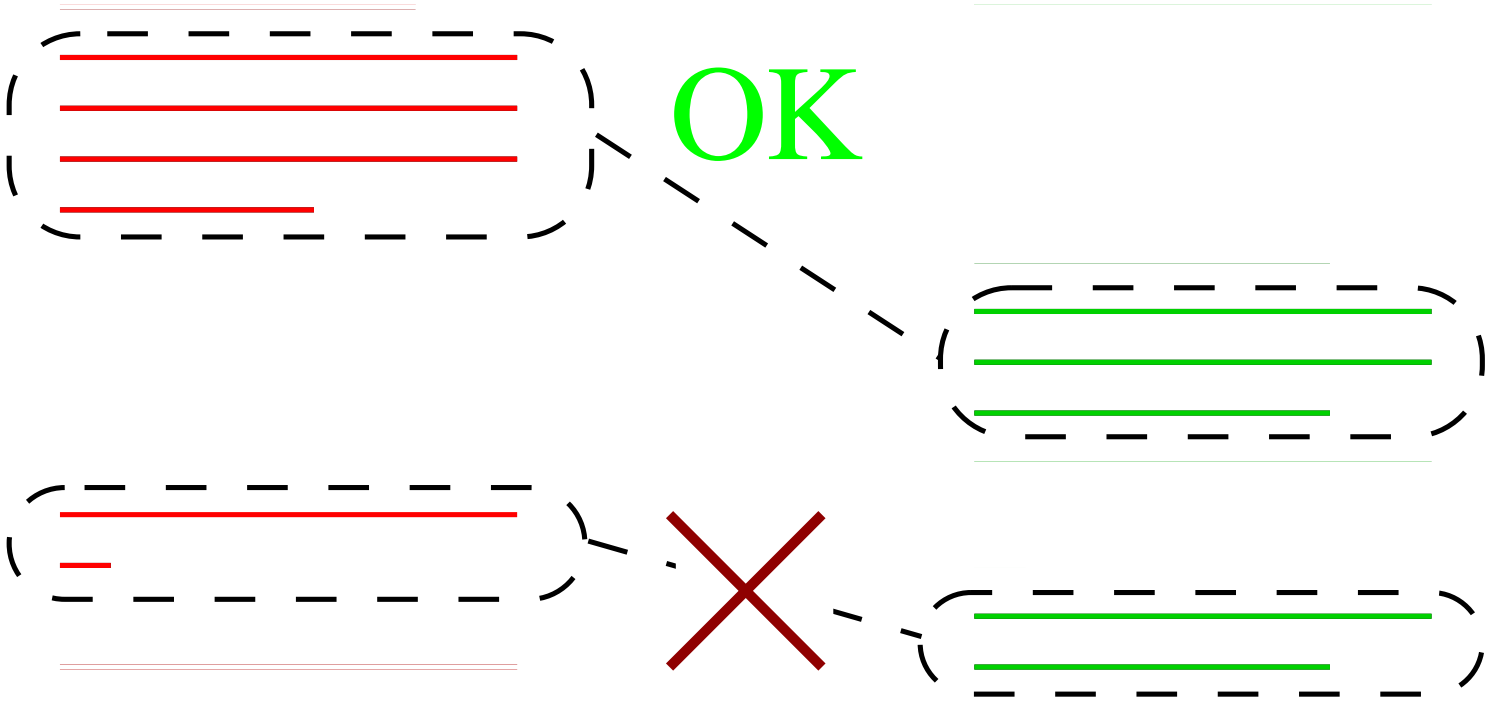


General Scheme: Pairing of Sentences



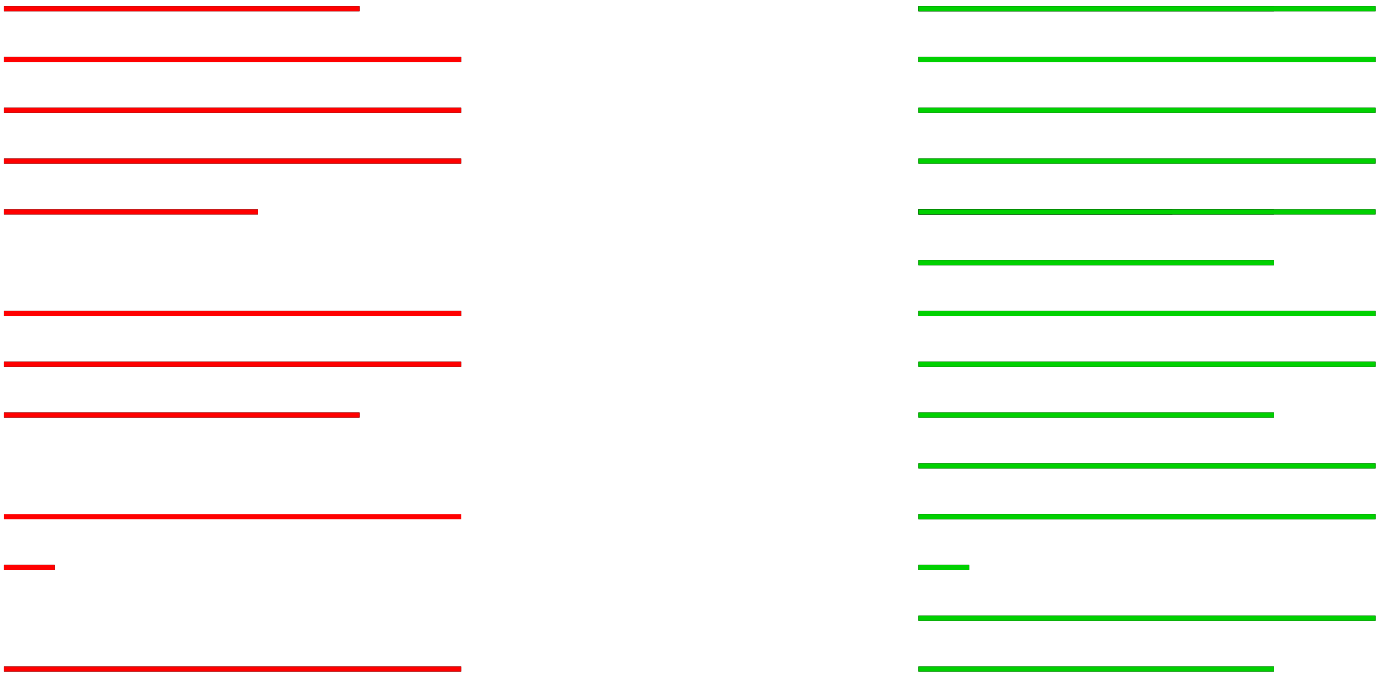
- by length, lexical overlap, IBM1, etc. (possibly, multi-step process)

General Scheme: Filtering Parallel Data

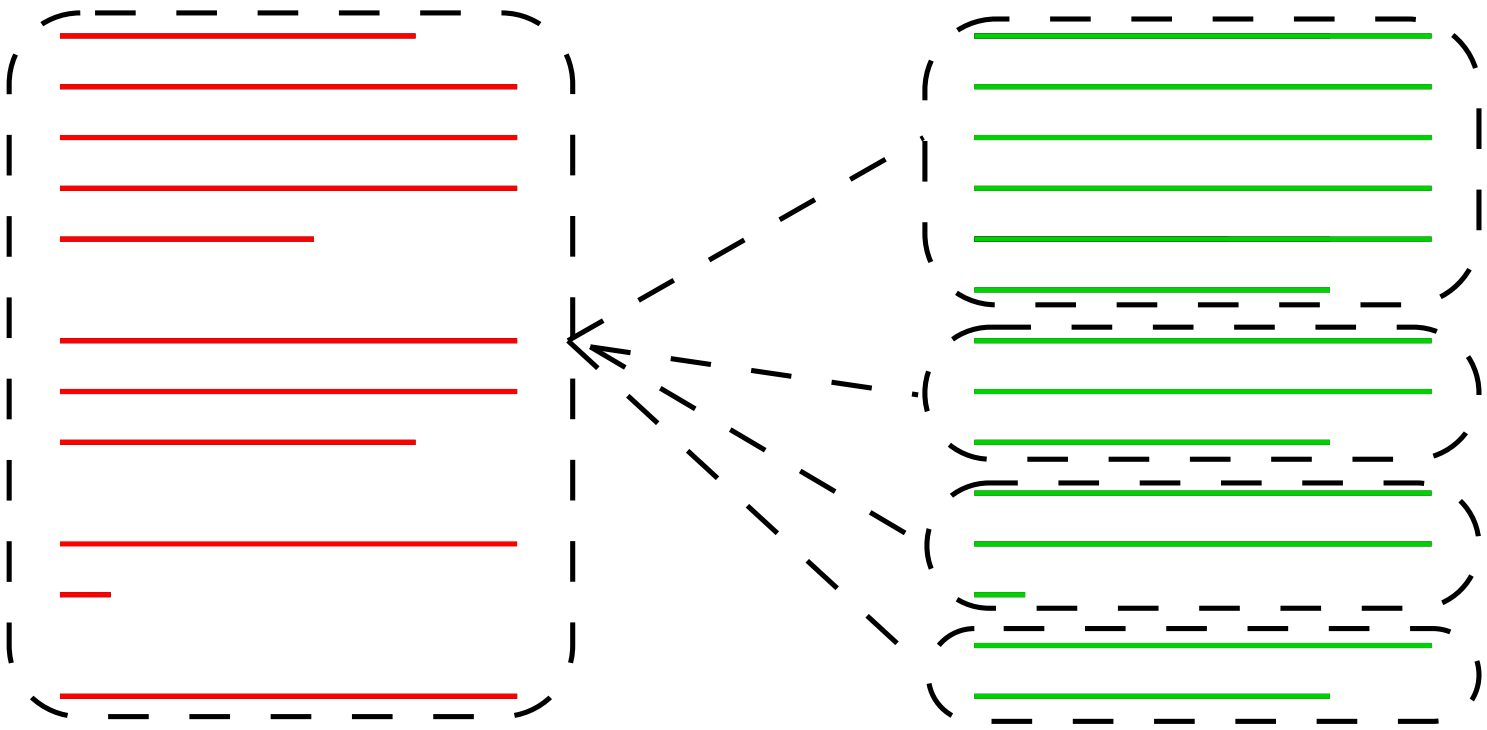


- by some threshold

Our Scheme: Pairing of Text



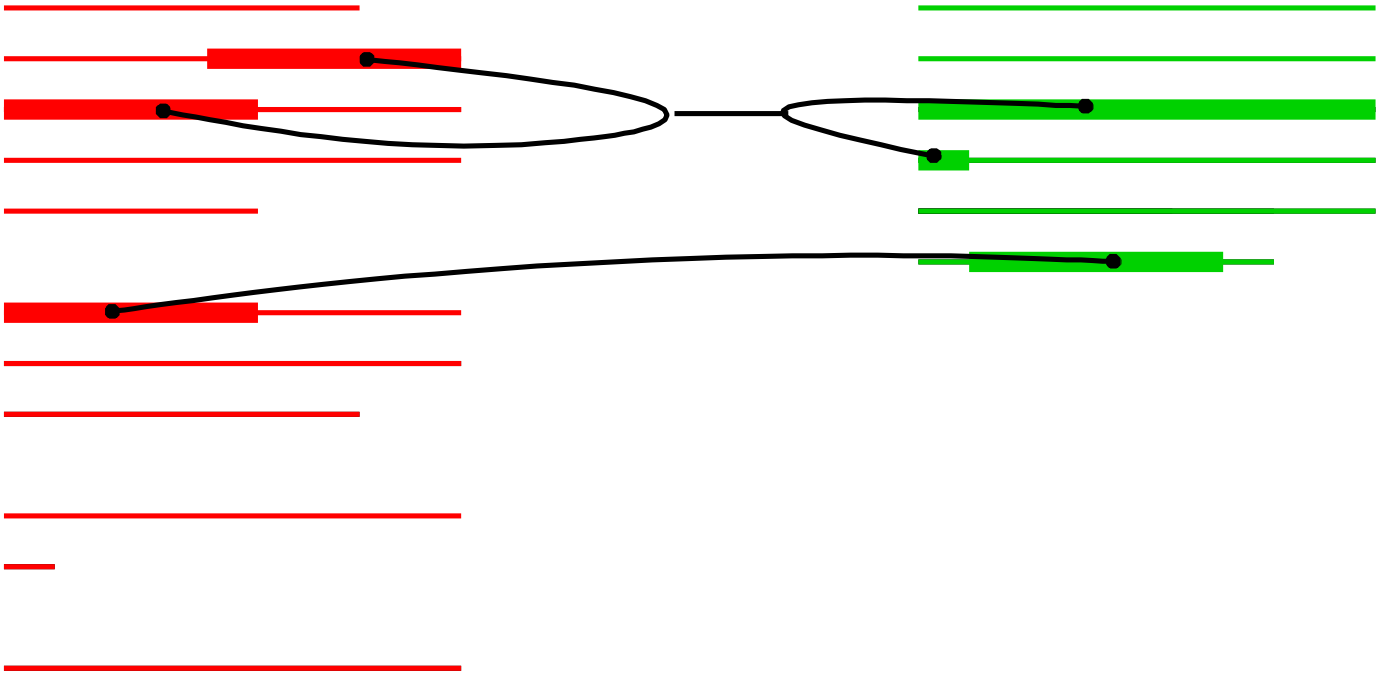
Our Scheme: Pairing of Text



- the whole source document is paired to each target sentence

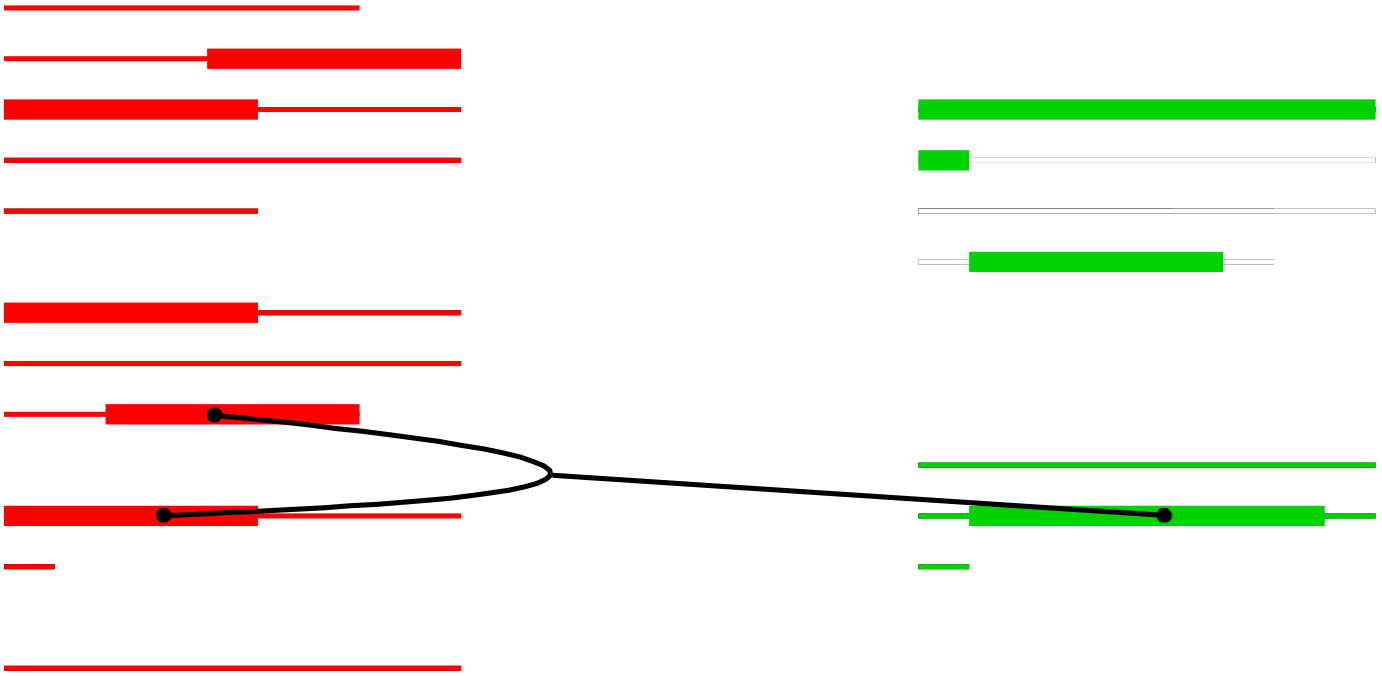
  aim: improve recall

Our Scheme: Filtering Parallel Data



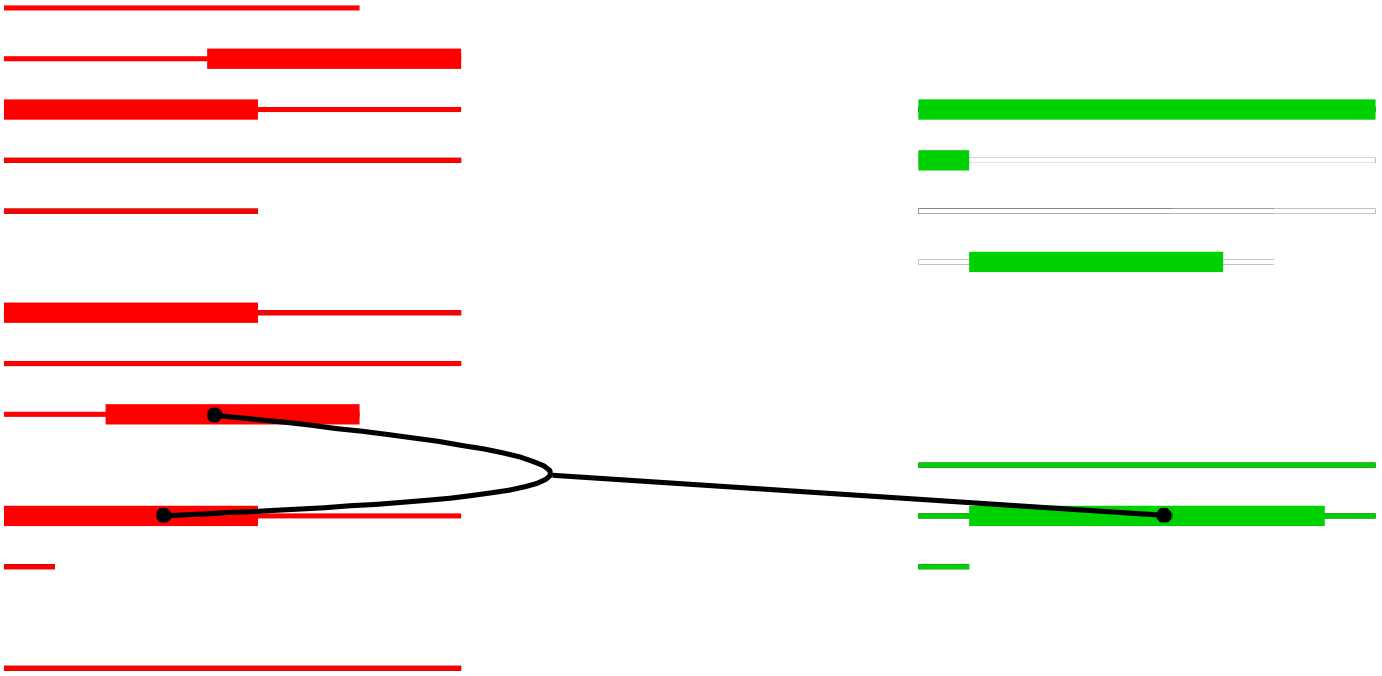
- looking for parallel **fragments** inside paired texts: not new in itself but ...

Our Scheme: Filtering Parallel Data



- looking for parallel **fragments** inside paired texts: not new in itself but ...
  - from one side, fragments across (source) sentences can be detected

Our Scheme: Filtering Parallel Data



- looking for parallel **fragments** inside paired texts: not new in itself but ...
  - from one side, fragments across (source) sentences can be detected
  - from the other...

A Novel Approach to Filtering Parallel Fragments

assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)

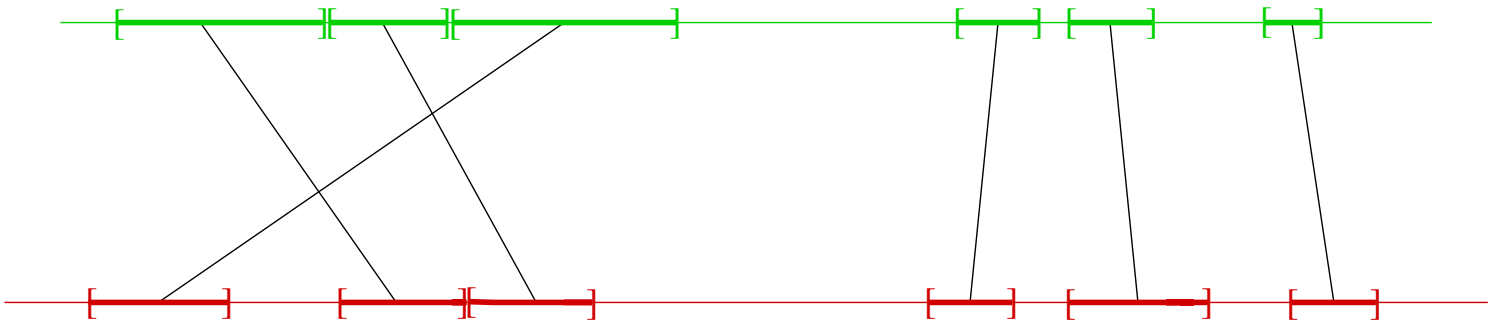
A Novel Approach to Filtering Parallel Fragments

assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)

1. For each text paired to be processed

A Novel Approach to Filtering Parallel Fragments

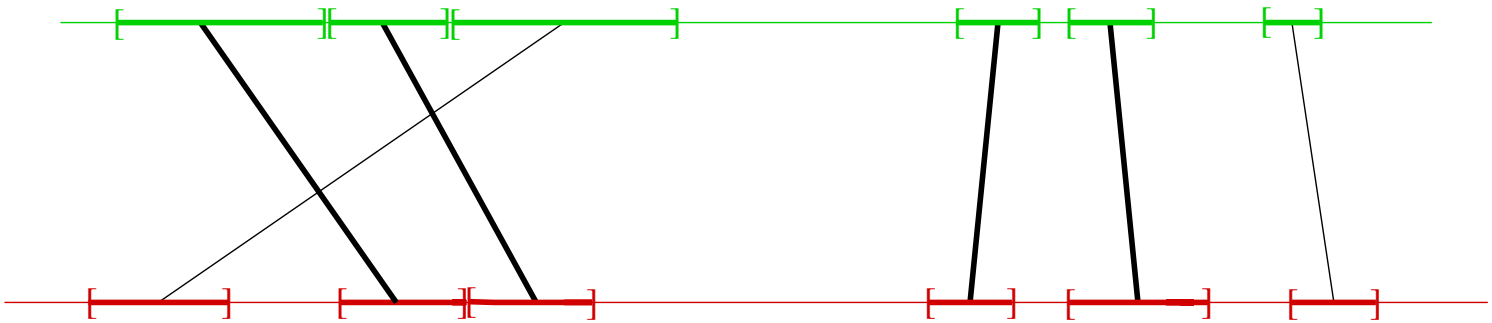
assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)



1. For each text paired to be processed
2. Compute partial phrase-level alignment through a constrained search

A Novel Approach to Filtering Parallel Fragments

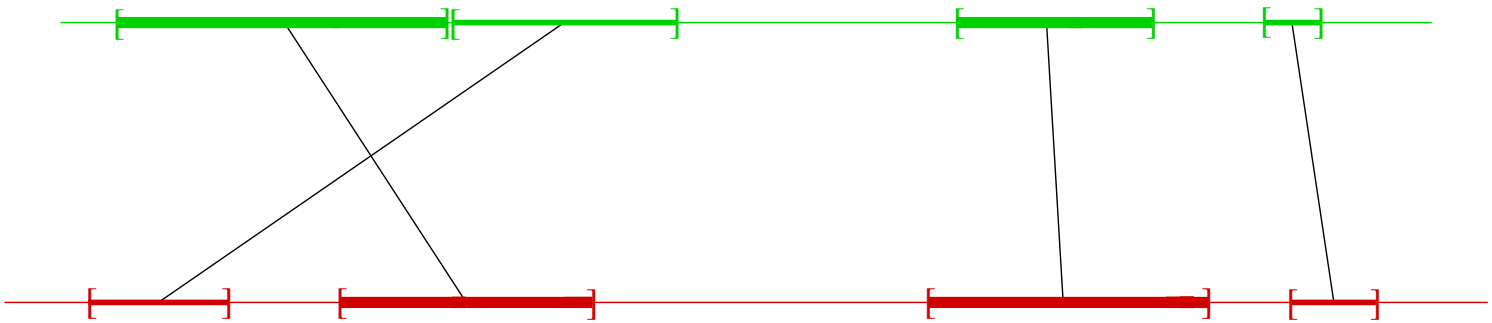
assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)



1. For each text paired to be processed
2. Compute partial phrase-level alignment through a constrained search
3. Extract fragments from the alignment by applying:
   (a) take the aligned phrases as parallel blocks
   (b) merge proximate blocks with non-aligned text in between
   (c) repeat (a-b) until no block pair can be merged

A Novel Approach to Filtering Parallel Fragments

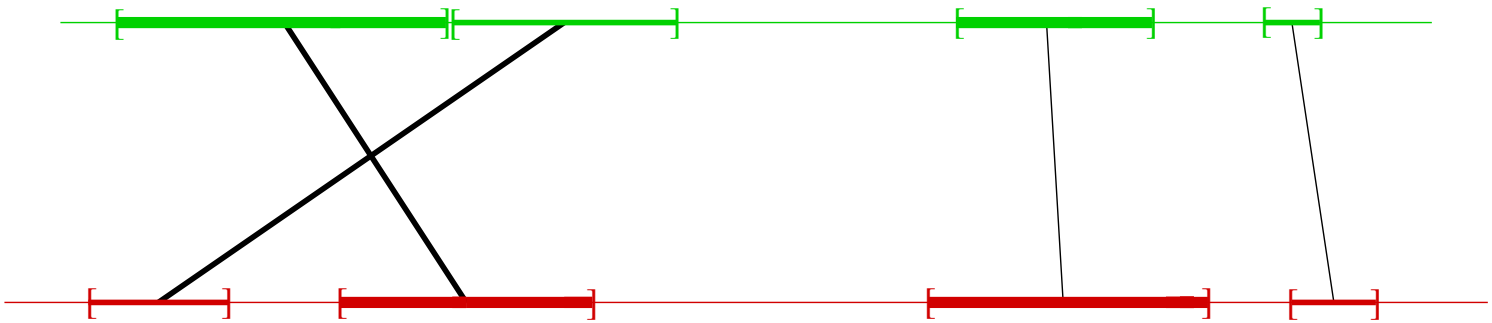
assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)



1. For each text paired to be processed
2. Compute partial phrase-level alignment through a constrained search
3. Extract fragments from the alignment by applying:
   (a) take the aligned phrases as parallel blocks
   (b) merge proximate blocks with non-aligned text in between
   (c) repeat (a-b) until no block pair can be merged

A Novel Approach to Filtering Parallel Fragments

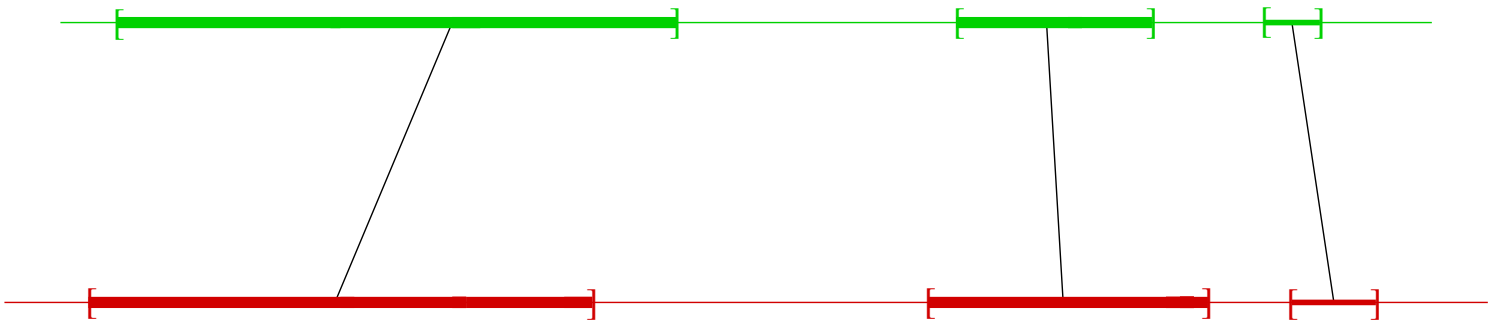
assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)



1. For each text paired to be processed
2. Compute partial phrase-level alignment through a constrained search
3. Extract fragments from the alignment by applying:
   (a) take the aligned phrases as parallel blocks
   (b) merge proximate blocks with non-aligned text in between
   (c) repeat (a-b) until no block pair can be merged

A Novel Approach to Filtering Parallel Fragments

assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)



1. For each text paired to be processed
2. Compute partial phrase-level alignment through a constrained search
3. Extract fragments from the alignment by applying:
   (a) take the aligned phrases as parallel blocks
   (b) merge proximate blocks with non-aligned text in between
   (c) repeat (a-b) until no block pair can be merged

A Novel Approach to Filtering Parallel Fragments

assumption: bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)

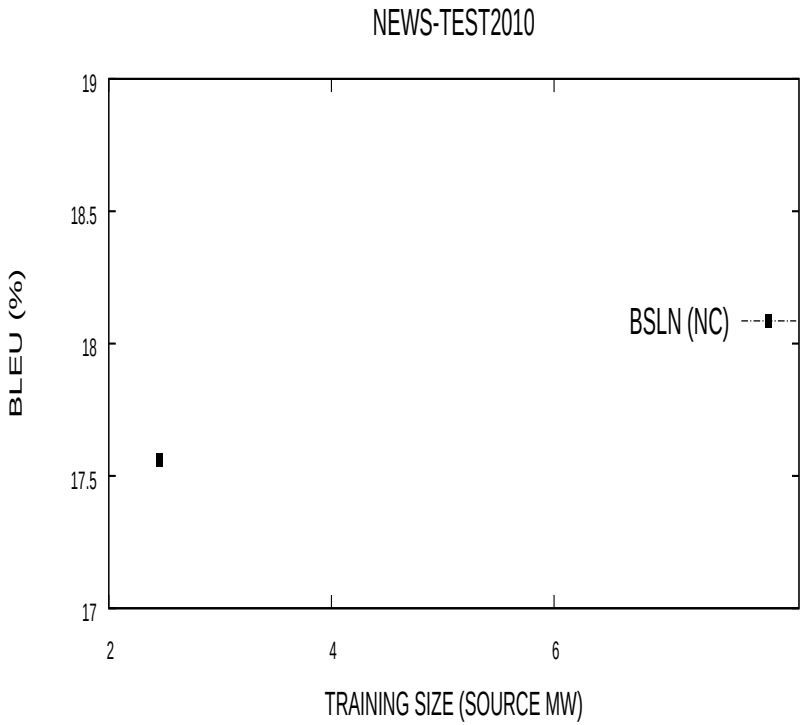
1. For each text paired to be processed
2. Compute partial phrase-level alignment through a constrained search
3. Extract fragments from the alignment by applying:
   (a) take the aligned phrases as parallel blocks
   (b) merge proximate blocks with non-aligned text in between
   (c) repeat (a-b) until no block pair can be merged
4. output final blocks as parallel fragments

A Novel Approach to Filtering Parallel Fragments

A few facts about our constrained search algorithm:

- no translation probabilities are used
- portions of either source or target texts can remain unaligned
- new words can be included by merging non-contiguous blocks
- it covers phrases rather than single words
- it is made efficient through DP, beam, pruning, ...

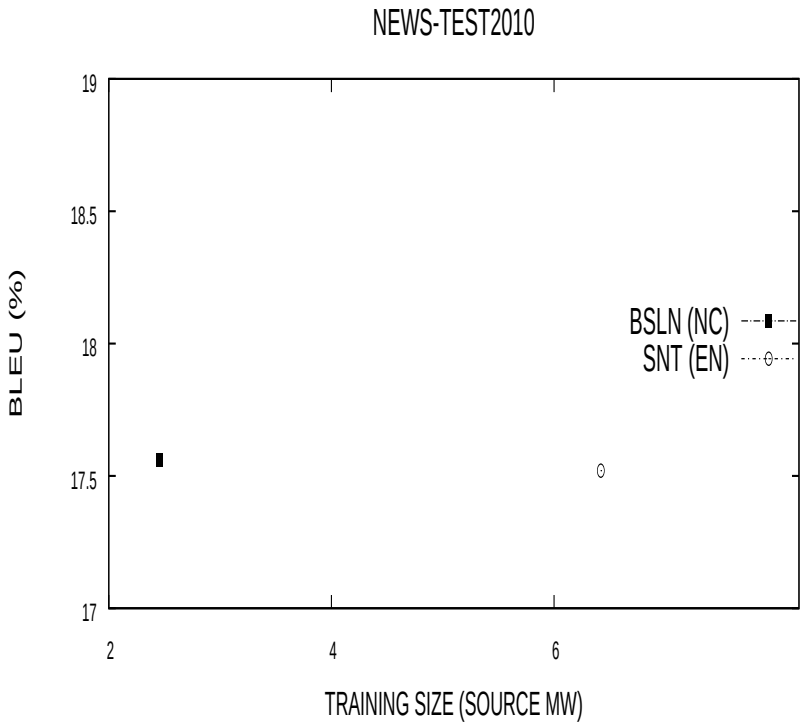
Evaluation



ACL WMT 2010 translation task: German news translated into English

NC: News Commentaries ($=$) EN: EuroNews ($\approx$ in-domain)

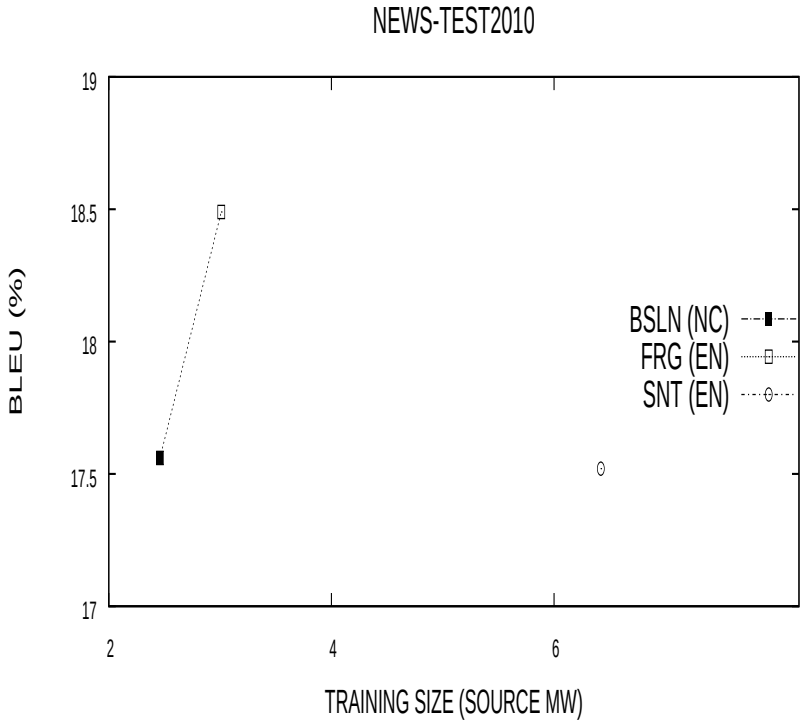
Evaluation



ACL WMT 2010 translation task: German news translated into English

NC: News Commentaries (=) EN: EuroNews (≈ in-domain)

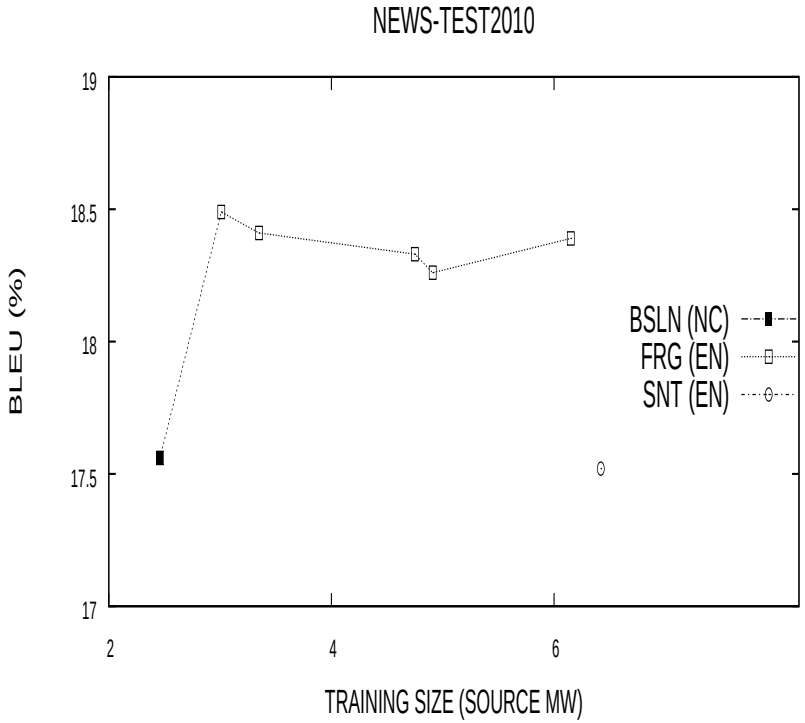
Evaluation



ACL WMT 2010 translation task: German news translated into English

NC: News Commentaries ($=$) EN: EuroNews ($\approx$ in-domain)

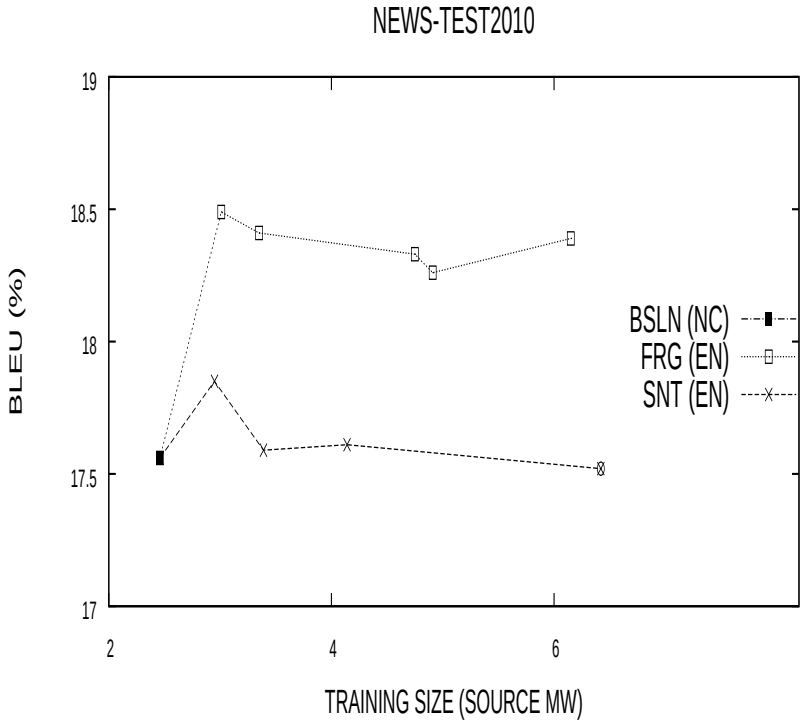
Evaluation



ACL WMT 2010 translation task: German news translated into English

NC: News Commentaries (=) EN: EuroNews (≈ in-domain)

Evaluation

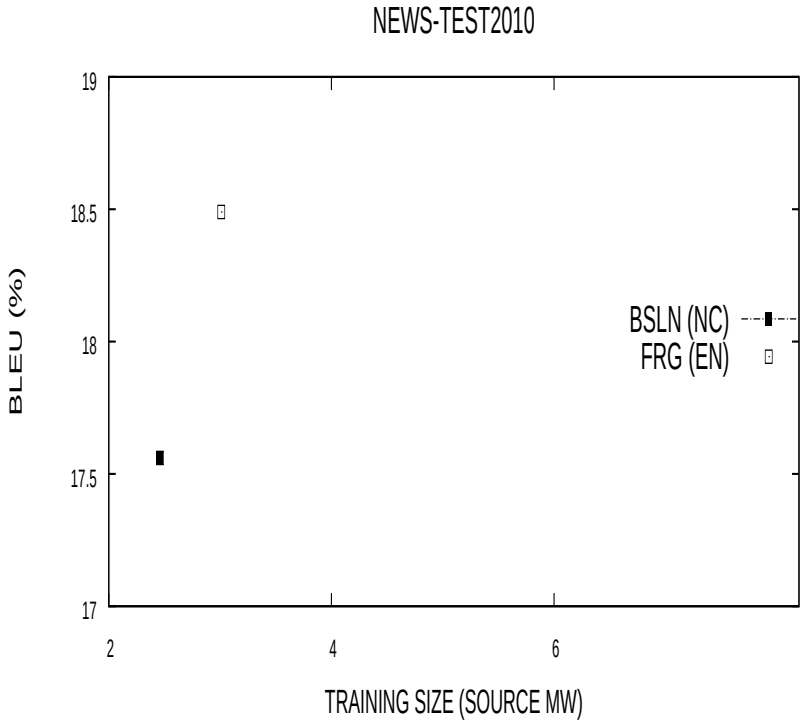


ACL WMT 2010 translation task: German news translated into English

NC: News Commentaries (=) EN: EuroNews (≈ in-domain)

SNT is a sentence-based filtering methods from the literature

Evaluation



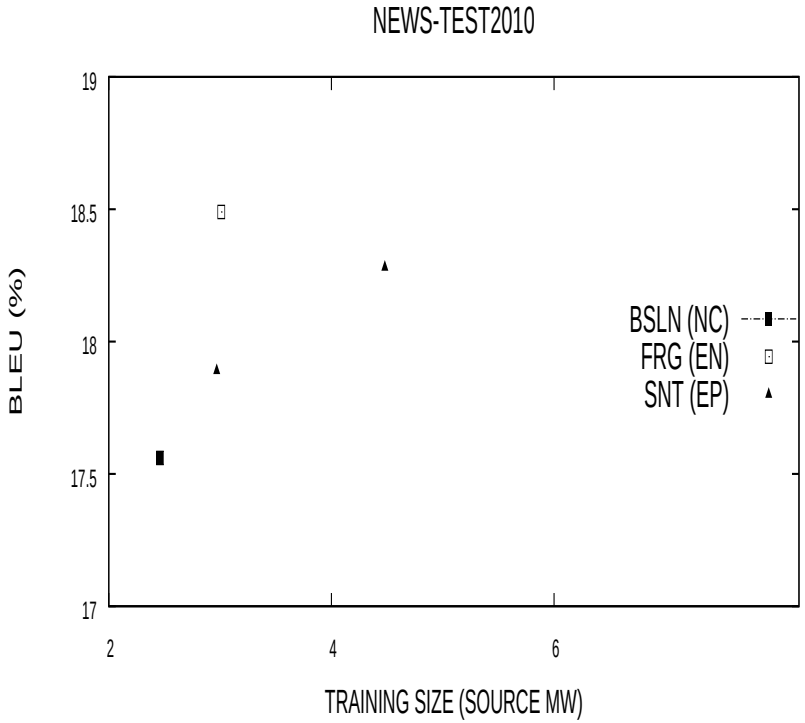
training data				NEWS TEST 2010
baseline		additional		
running words (source)	type	running words (source)	type	
2.5M	NC	-	-	17.56
2.5M	NC	0.5M	FRG(EN)	18.49

ACL WMT 2010 translation task: German news translated into English

NC (=) EN (≈ in-domain) EP: EuroParl (= out-of-domain)

SNT is a sentence-based filtering methods from the literature

Evaluation



training data				NEWS TEST 2010
baseline		additional		
running words (source)	type	running words (source)	type	
2.5M	NC	-	-	17.56
2.5M	NC	0.5M	FRG(EN)	18.49
2.5M	NC	0.5M	SNT(EP)	17.89
2.5M	NC	2.0M	SNT(EP)	18.28

ACL WMT 2010 translation task: German news translated into English

NC (=) EN (≈ in-domain) EP: EuroParl (= out-of-domain)

SNT is a sentence-based filtering methods from the literature

Conclusions

- Applying standard training on in-domain comparable data is not effective
- Better mining parallel fragments within comparable documents
  - fragments are preferable to whole sentences (less noisy)
- Experimental results show that the method is effective:
  - BLEU score increases up to 5% relative
  - equivalent to using four-times more out-of-domain parallel data

Thanks