

# 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

## Problem

- SMT needs sufficient translated text as training material
- large parallel corpora **do not exist** for many language pairs
- → 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

Sie sehen neben schärferen Überwachungsmöglichkeiten auch die Schaffung eines Muslimrates vor, um Extremisten schneller aufspüren zu können.

<sup>1</sup>[Der italienische Innenminister Giuseppe Pisanu sagte, dass] <sup>1</sup>dieser Rat <sup>2</sup>[eine Art italienischen Islam schaffen soll, der die nationale Identität und die Gesetze respektiert.

Gleichzeitig] <sup>2</sup> sollen die islamische Identität und ihre Andersartigkeit geschützt werden, solange sie staats-treu sind.

Dem Gesetzeswerk soll nun noch das italienische Unterhaus vor der Sommerpause zustimmen.

Hamza Roberto Piccardo von der Union der muslimischen Gemeinden in Italien kritisierte die Schaffung eines neuen Muslimrates als nutzlos, um die Gefahr von Anschlägen zu senken.

<sup>3</sup>[Andere italienische Muslime meinen aber, es] <sup>3</sup> sei eine gute Initiative, <sup>4</sup>[da die Menschen sonst alle Glaubensgenossen mit den Extremisten] <sup>4</sup> und Attentätern in einen Topf werfen.

leaders.

<sup>1</sup>[Interior Minister Giuseppe Pisanu said:] <sup>1</sup>"The council will move towards <sup>2</sup>[the creation of an Italian Islam, respectful of our national identity and our laws and at the same time] <sup>2</sup> protected in its identity."

The aim of the body is to give advice on the new security legislation and open a channel of communication with Muslims.

But some believe the government is over-reacting to a perceived terrorist threat.

One Islamic community leader said: "If the government creates a council for security reasons I think it will not counter the terror threat."

<sup>3</sup>[But other Muslims have] <sup>3</sup> welcomed the initiative.

"I believe it's a good idea <sup>4</sup>[because people tend to associate all Muslims with extremists] <sup>4</sup> who are responsible for attacks," said one man .

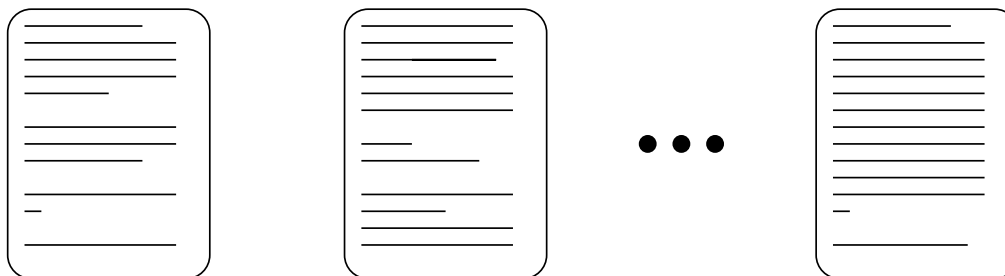
**Previous work** on how to collect parallel data from comparable data:

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

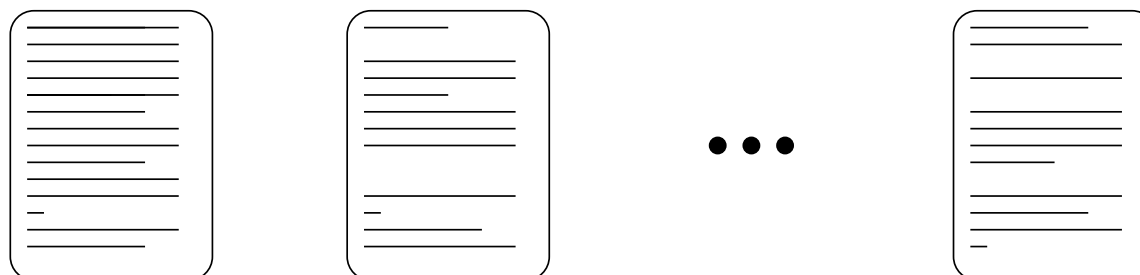
**Our approach** fits this scheme and it is new on some aspects.

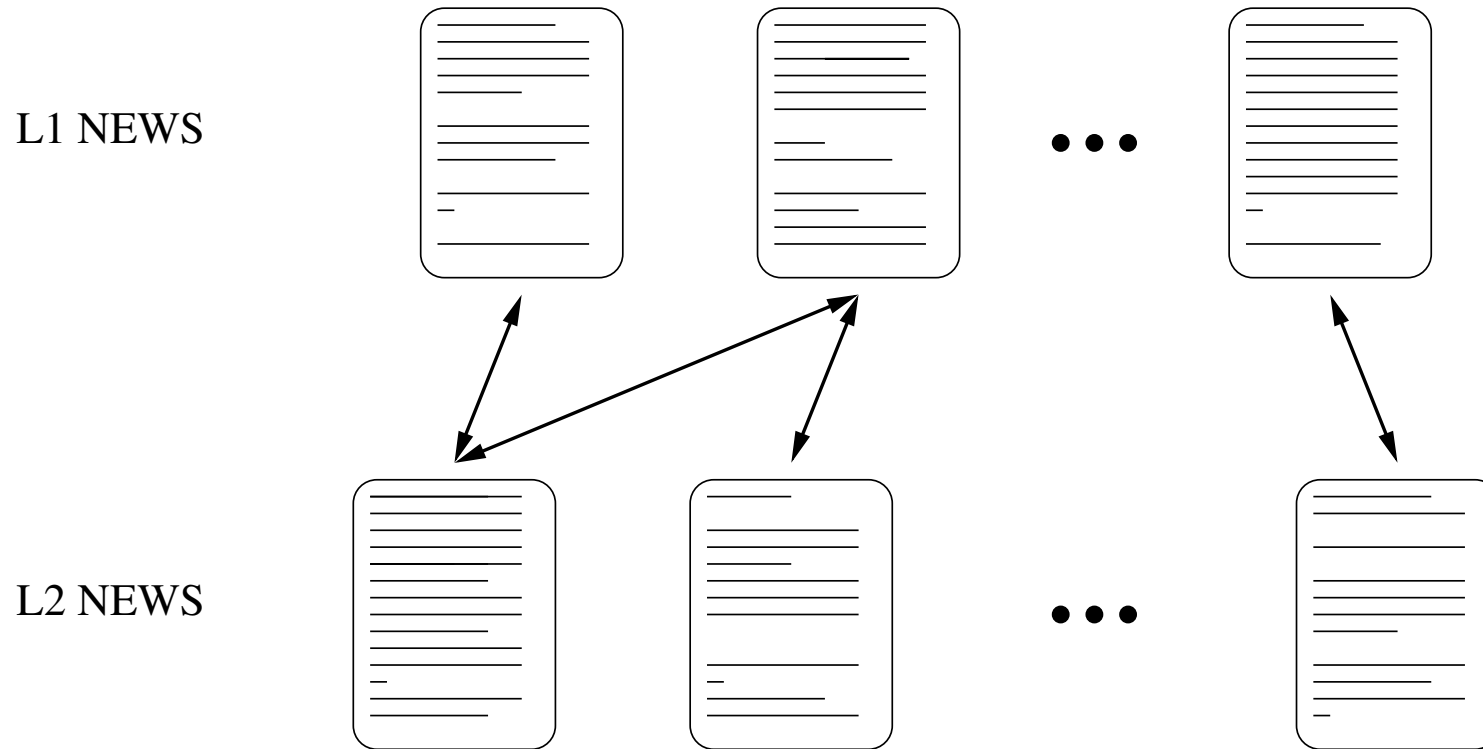
# General Scheme

L1 NEWS



L2 NEWS

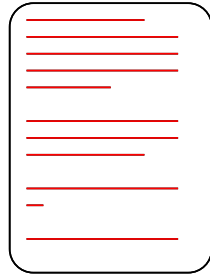




- by exploiting metadata (images, explicit links)

# General Scheme: Sentence Splitting

L1 NEWS

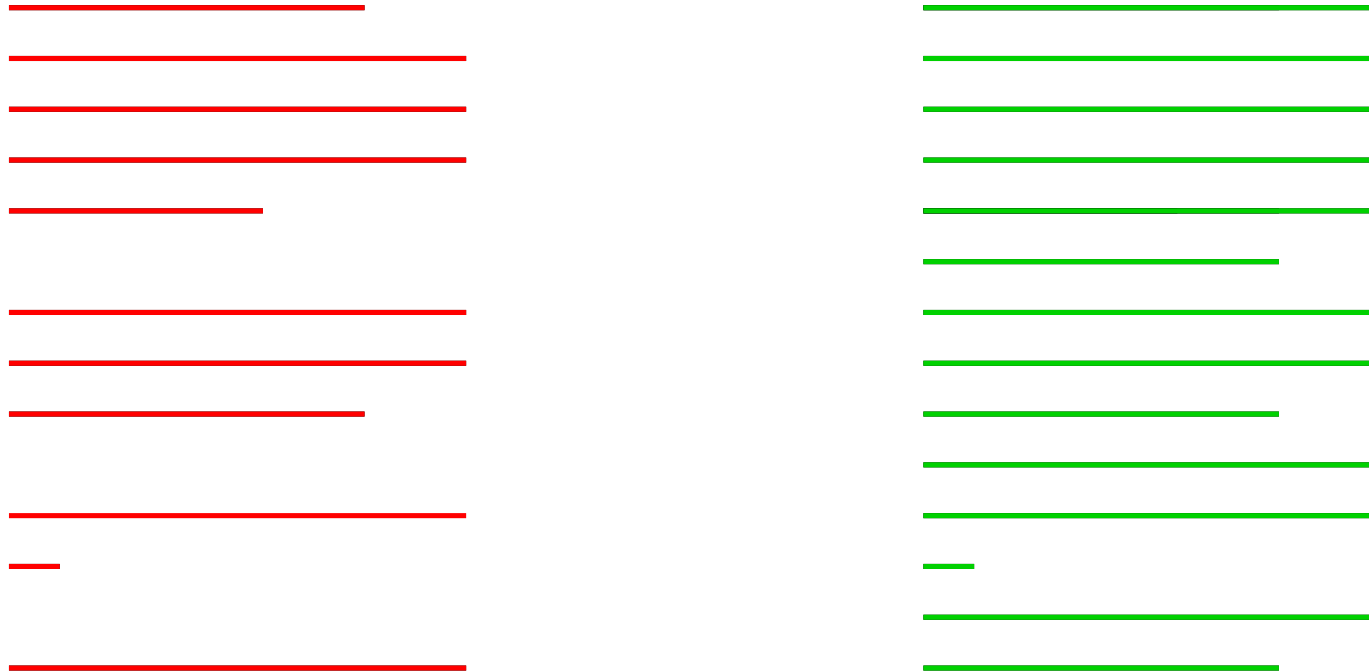


L2 NEWS



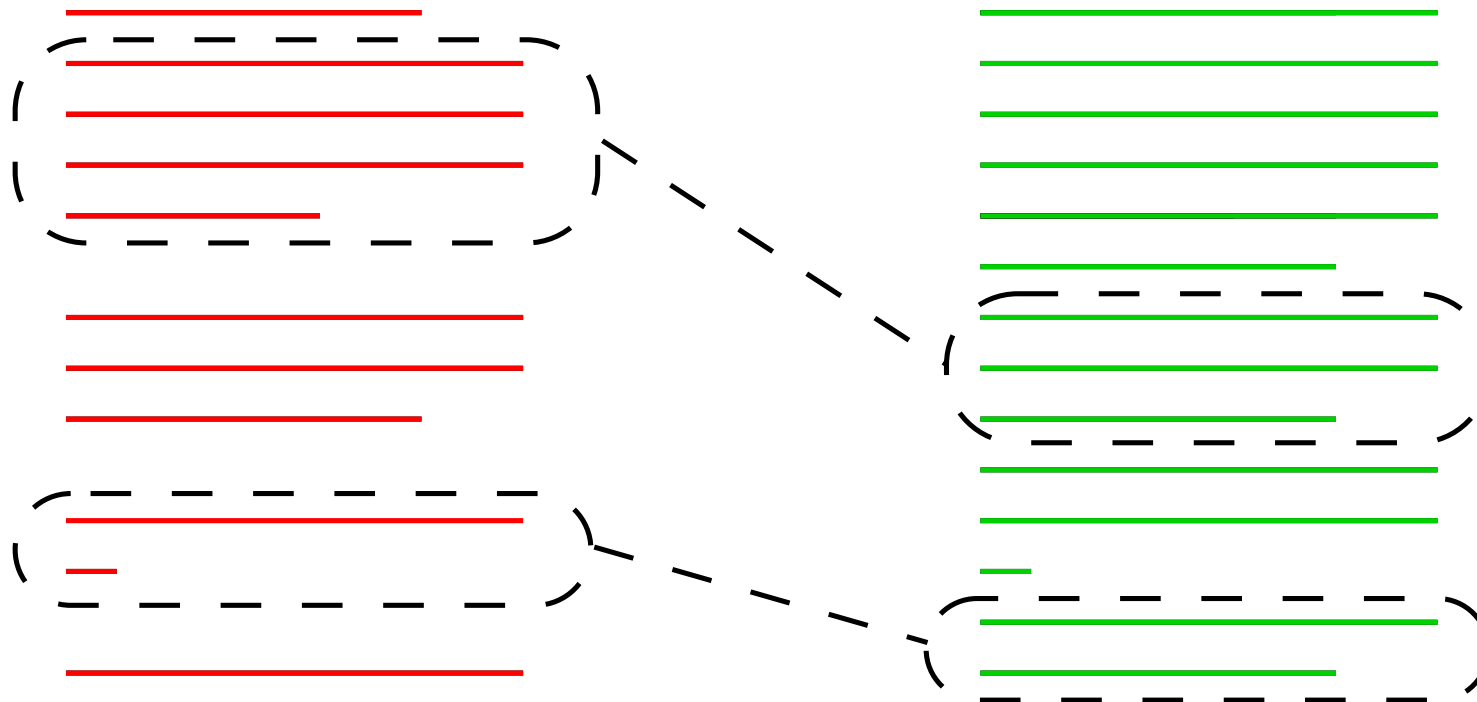
- on strong punctuation

# General Scheme: Pairing of Sentences



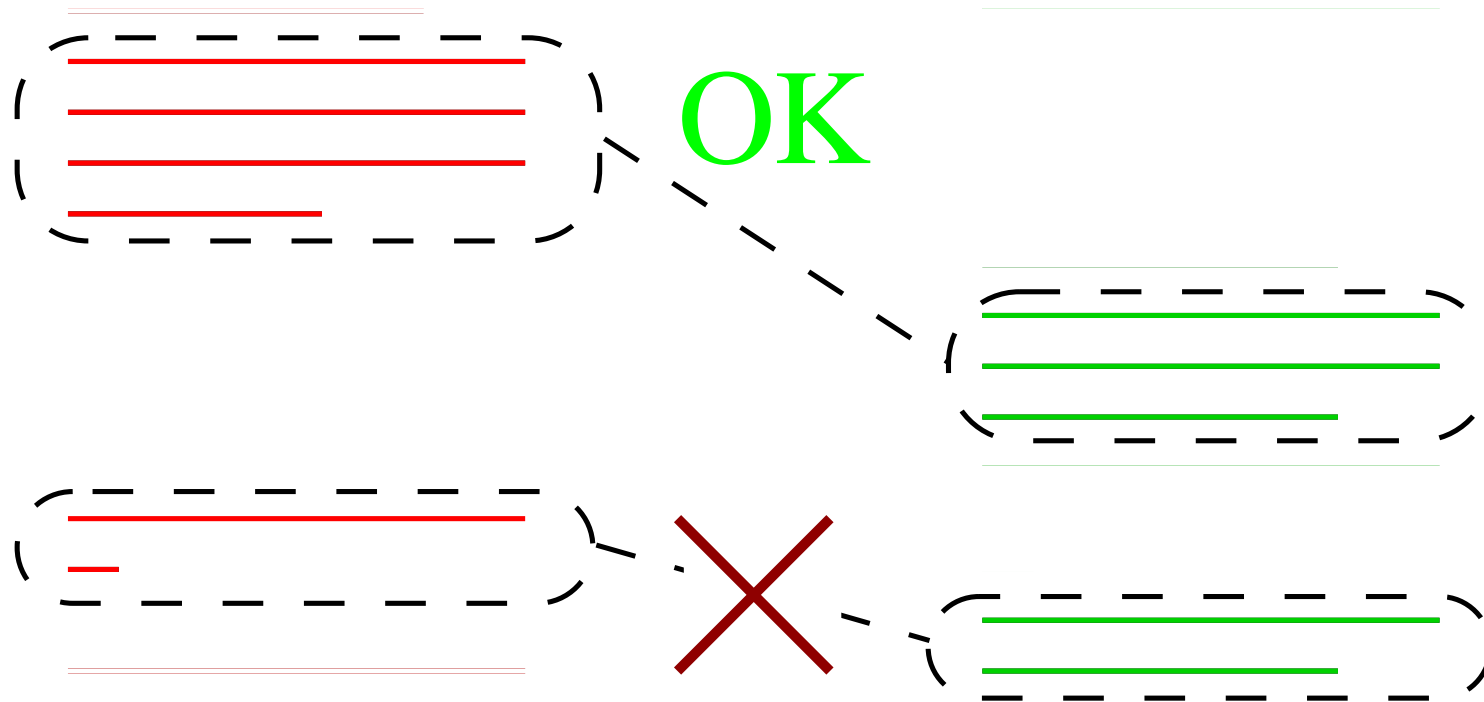


# General Scheme: Pairing of Sentences



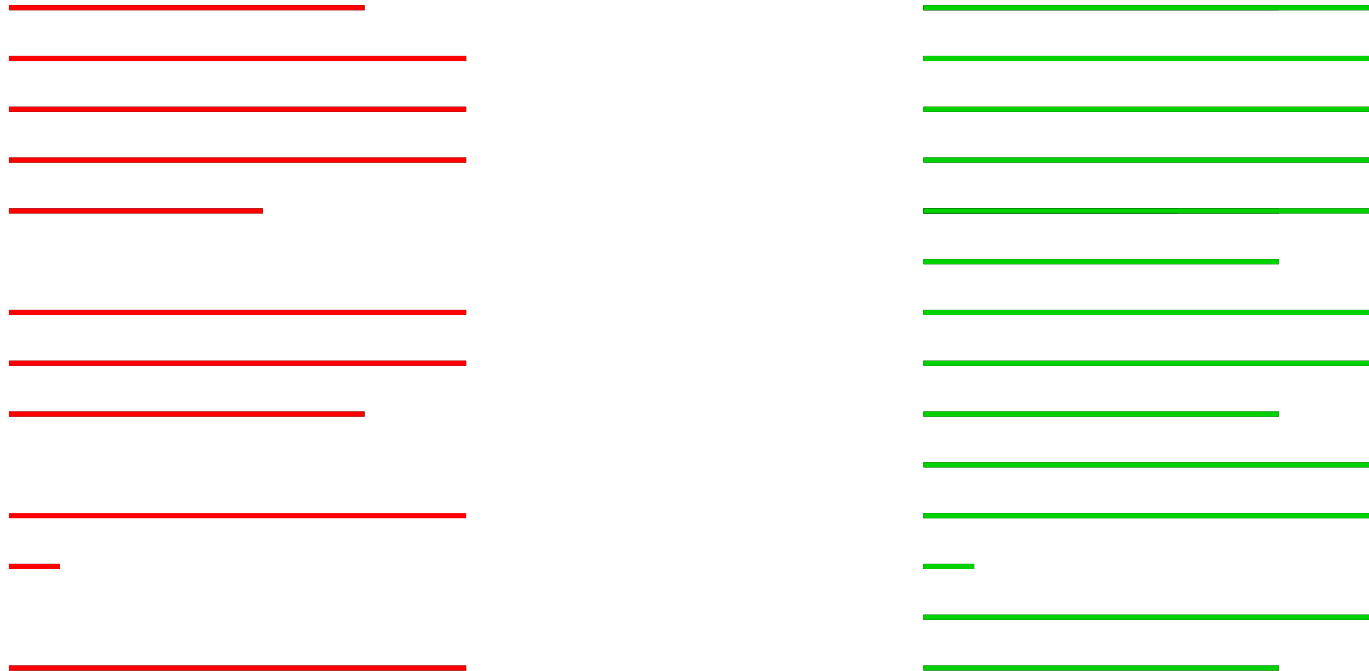
- by length, lexical overlap, IBM1, etc. (possibly, multi-step process)

# General Scheme: Filtering Parallel Data

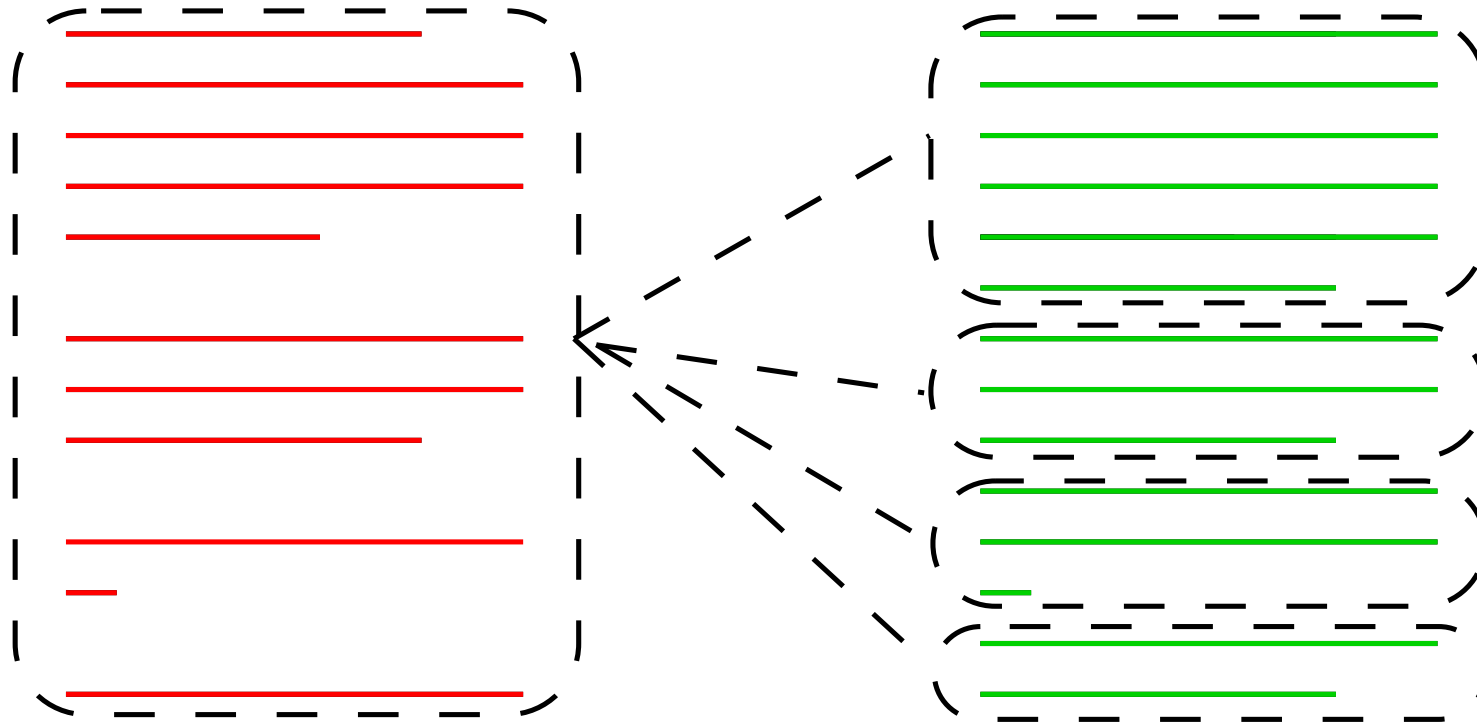


- by some threshold

# Our Scheme: Pairing of Text



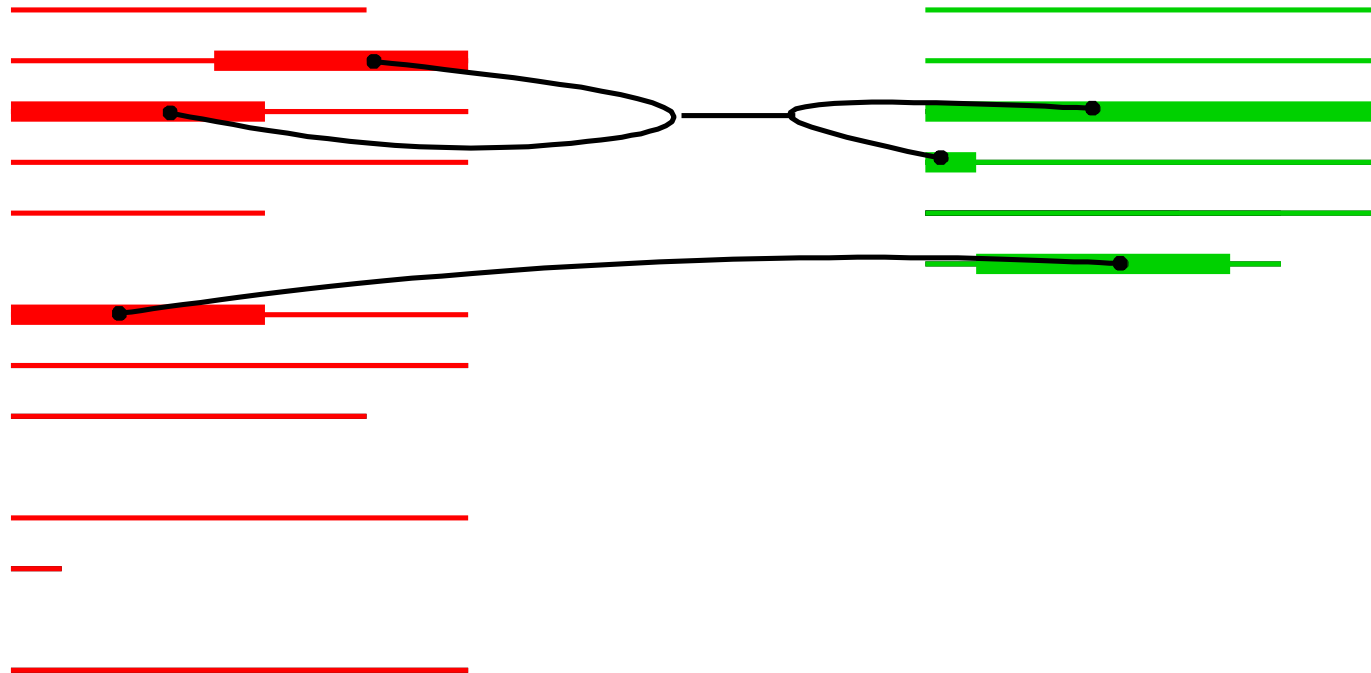
# Our Scheme: Pairing of Text



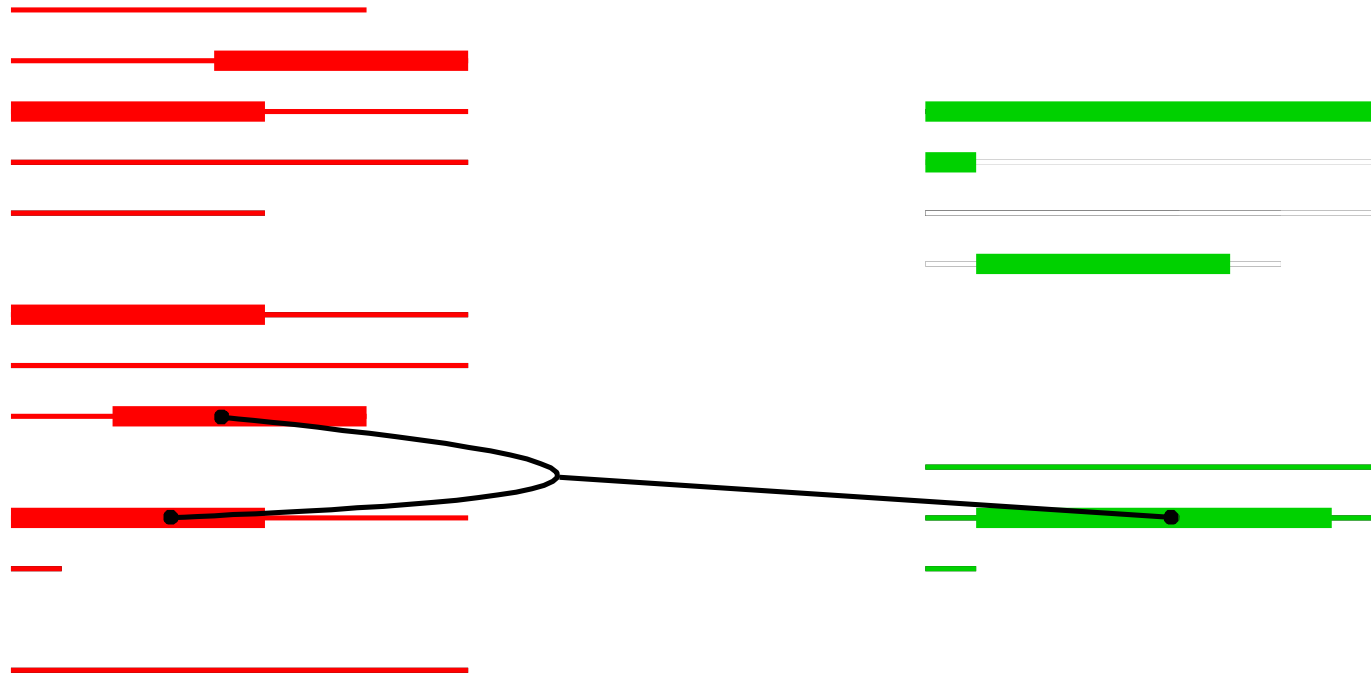
- the whole source document is paired to each target sentence

aim: improve recall

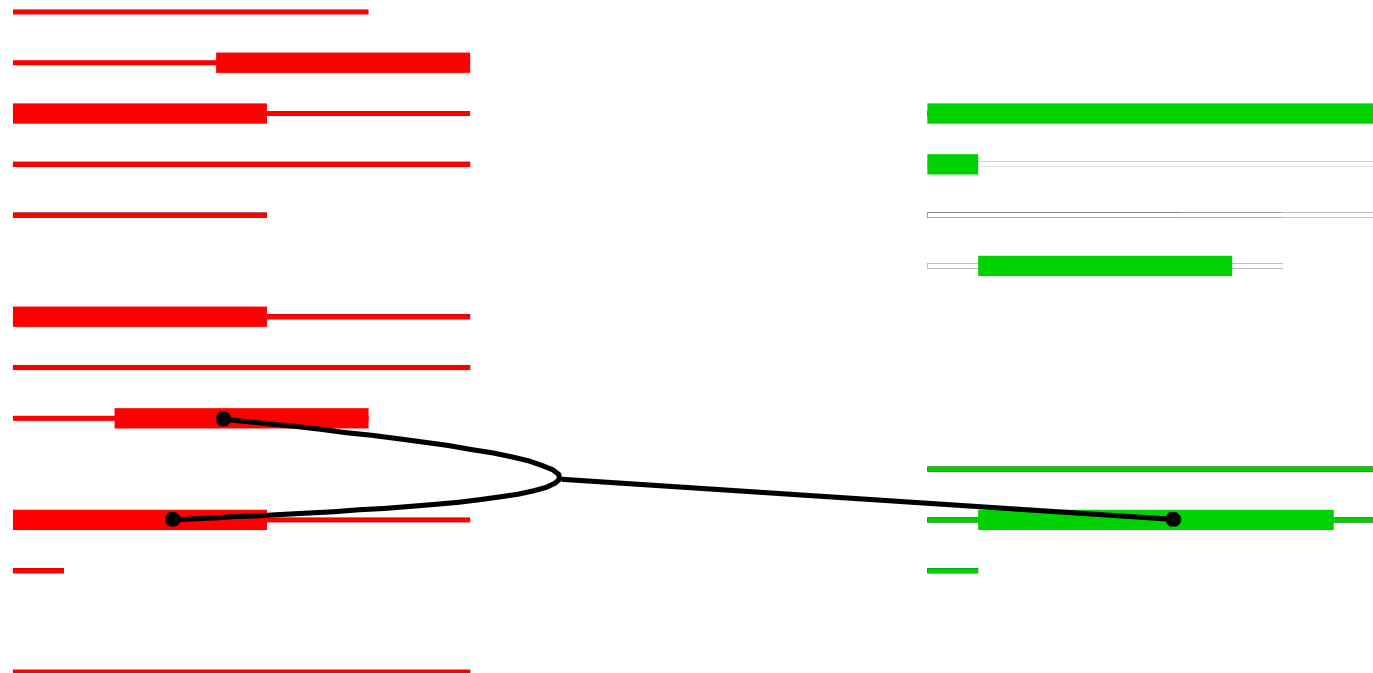
# Our Scheme: Filtering Parallel Data



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



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



- 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...

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



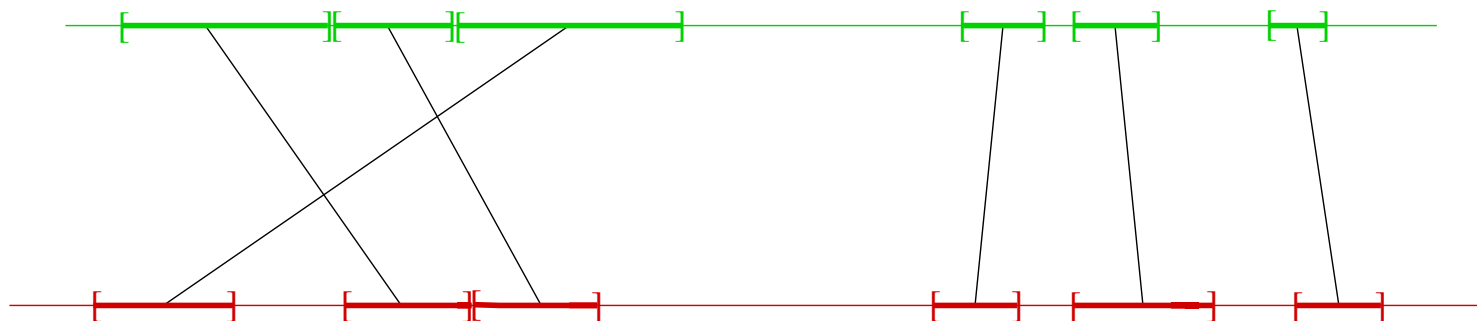
**assumption:** bootstrap from a phrase-pair repository (trans. mem., trans. mdl...)

---

---

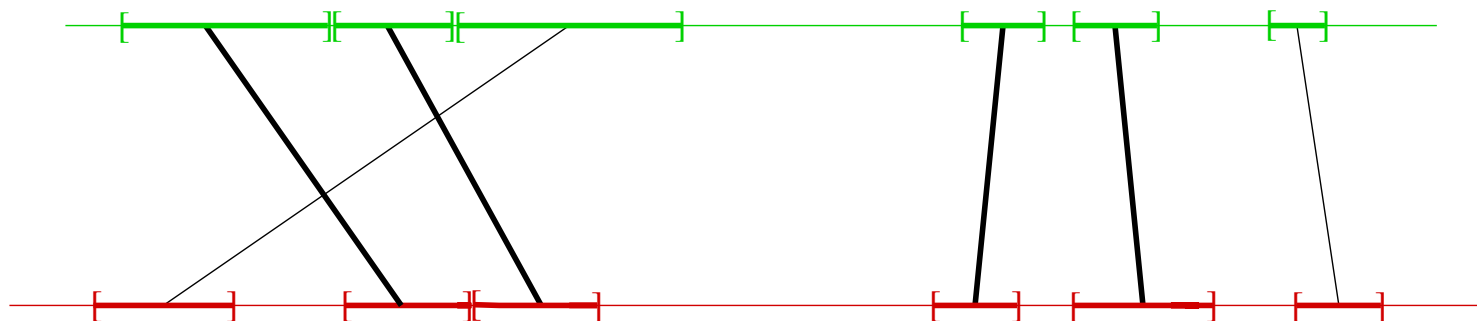
1. For each text paired to be processed

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



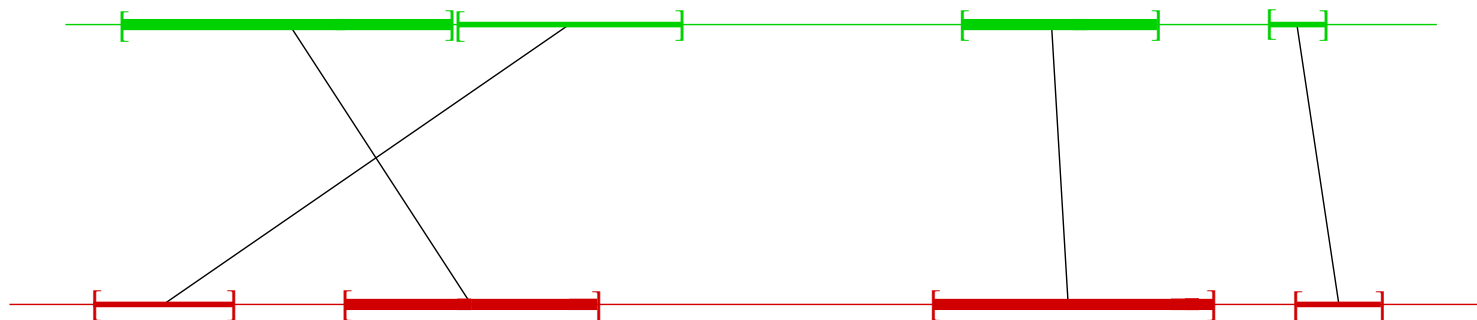
1. For each text paired to be processed
2. Compute partial phrase-level alignment through a constrained search

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



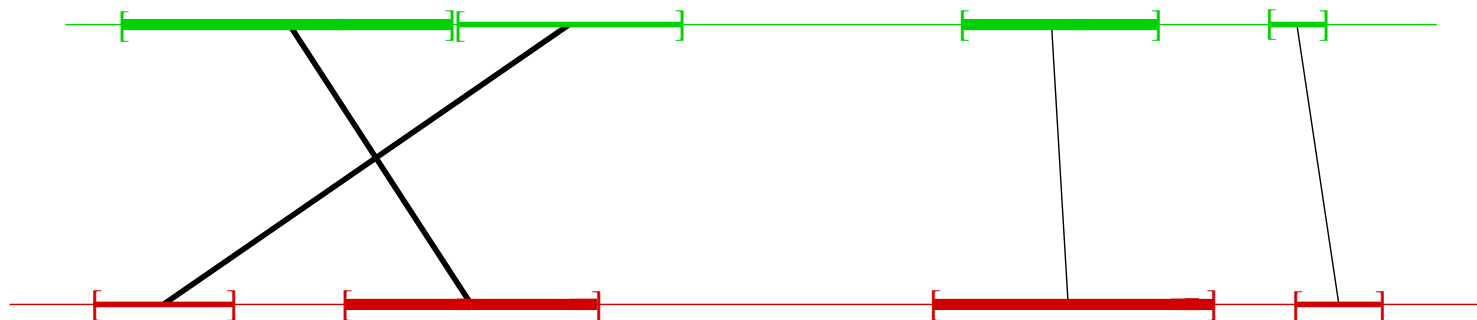
1. For each text paired to be processed
2. Compute partial [phrase-level alignment](#) through a [constrained search](#)
3. Extract fragments from the alignment by applying:
  - (a) take the aligned phrases as parallel blocks
  - (b) merge proximate blocks with non-aligned text in between
  - (c) repeat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

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



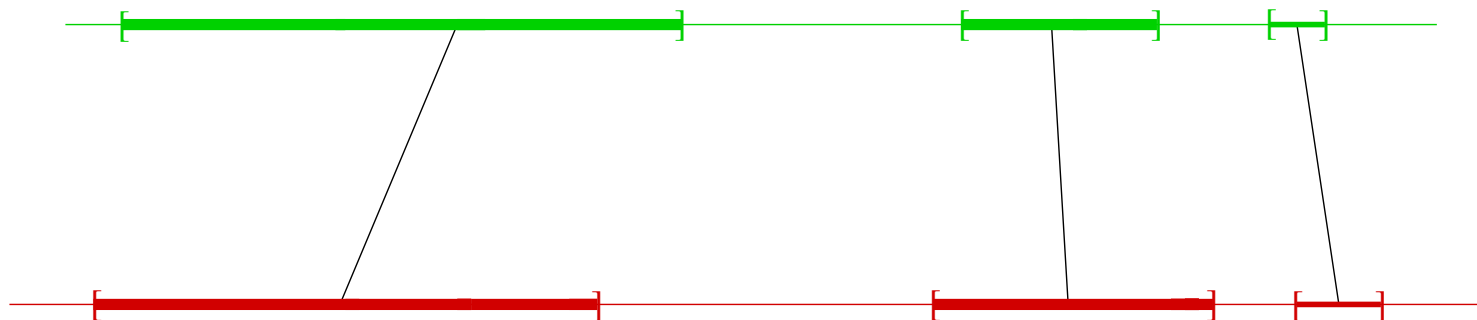
1. For each text paired to be processed
2. Compute partial [phrase-level alignment](#) through a [constrained search](#)
3. Extract fragments from the alignment by applying:
  - (a) take the aligned phrases as parallel blocks
  - (b) merge proximate blocks with non-aligned text in between
  - (c) repeat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

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



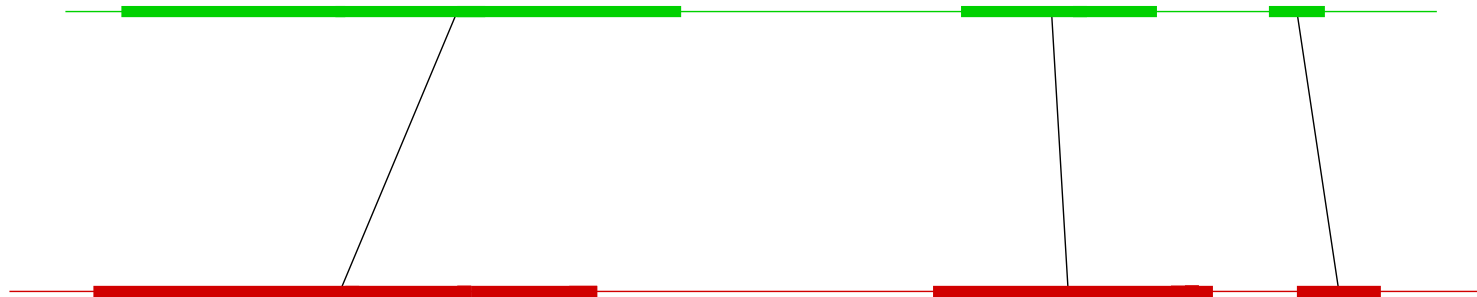
1. For each text paired to be processed
2. Compute partial [phrase-level alignment](#) through a [constrained search](#)
3. Extract fragments from the alignment by applying:
  - (a) take the aligned phrases as parallel blocks
  - (b) merge proximate blocks with non-aligned text in between
  - (c) repeat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

**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

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

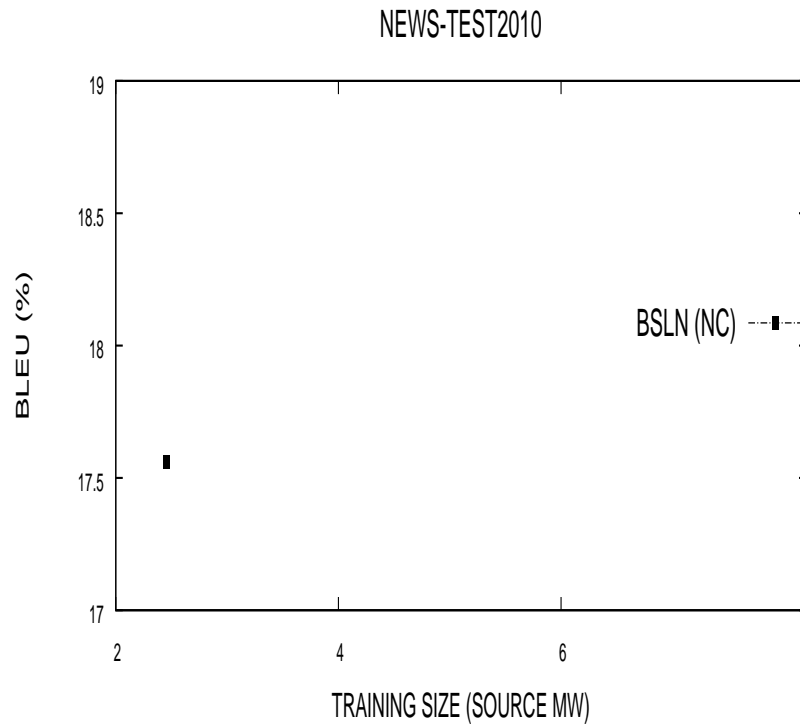


1. For each text paired to be processed
2. Compute partial [phrase-level alignment](#) through a [constrained search](#)
3. Extract fragments from the alignment by applying:
  - (a) take the aligned phrases as parallel blocks
  - (b) merge proximate blocks with non-aligned text in between
  - (c) repeat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

A few facts about our constrained search algorithm:

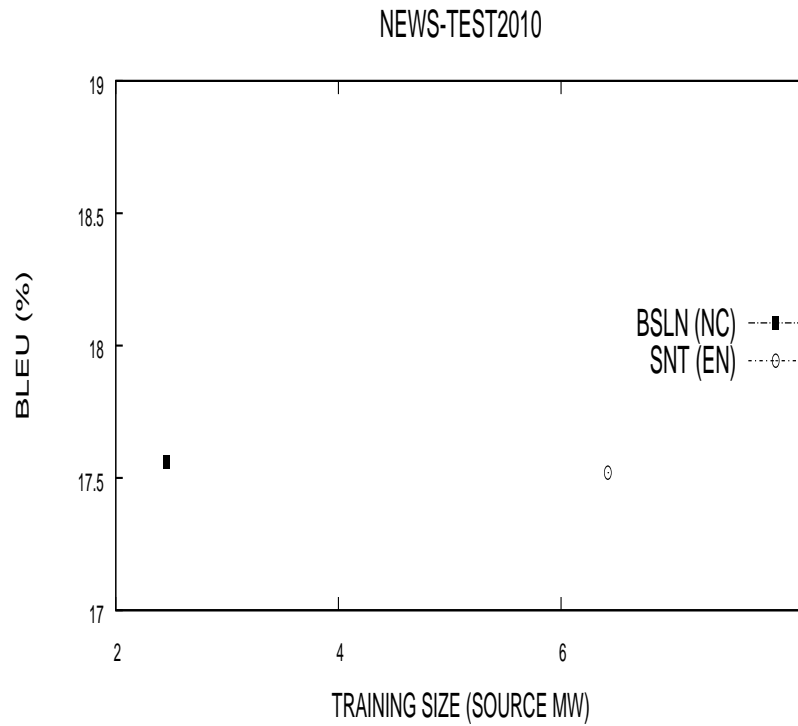
- no translation probabilities are used
- portions of either source or target texts can remain unaligned
- new words can be included by merging non-contiguous blocks
- it covers phrases rather than single words
- it is made efficient through DP, beam, pruning, ...





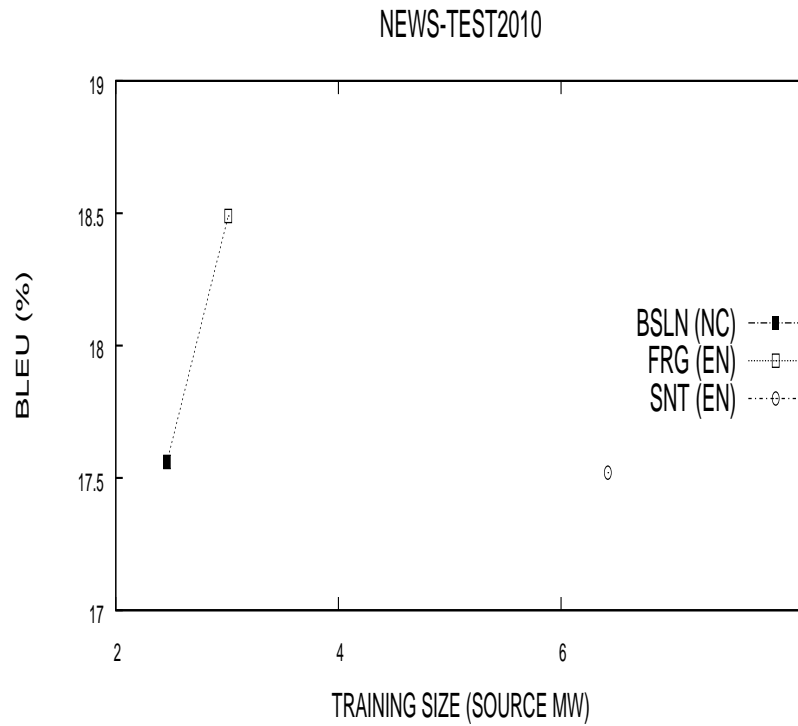
ACL WMT 2010 translation task: German news translated into English

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



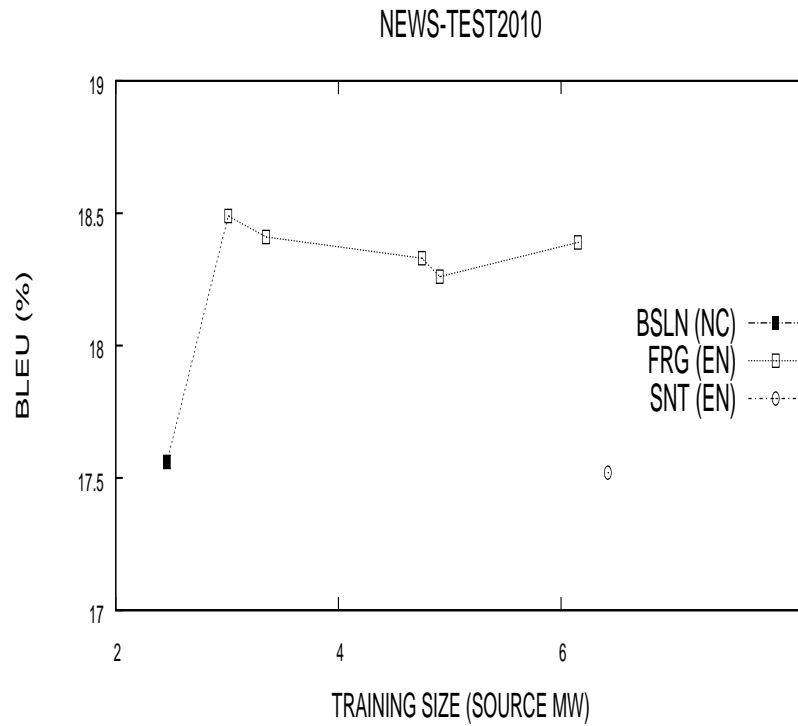
ACL WMT 2010 translation task: German news translated into English

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



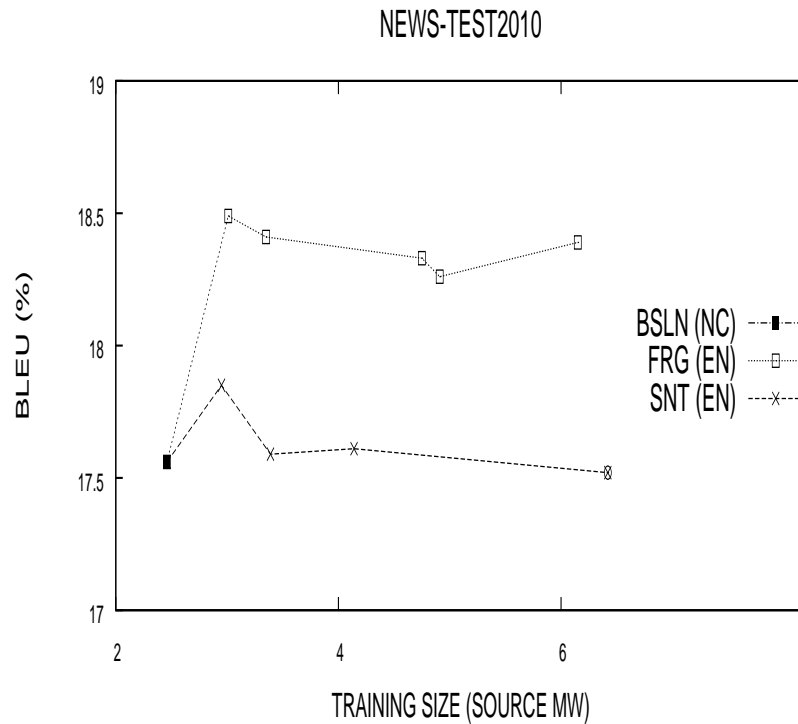
ACL WMT 2010 translation task: German news translated into English

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



ACL WMT 2010 translation task: German news translated into English

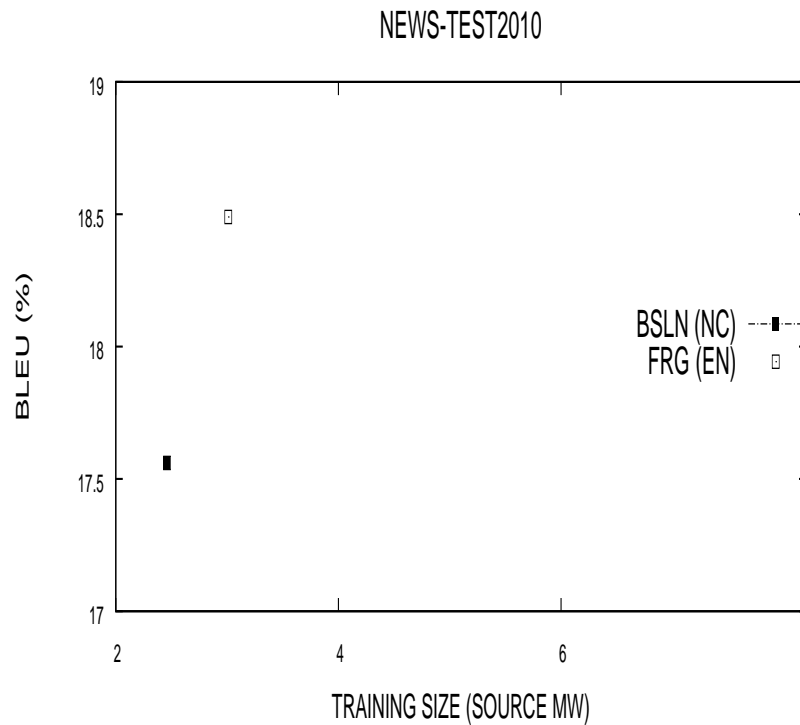
NC: News Commentaries (=)      EN: EuroNews ( $\approx$  in-domain)



ACL WMT 2010 translation task: German news translated into English

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

SNT is a sentence-based filtering methods from the literature

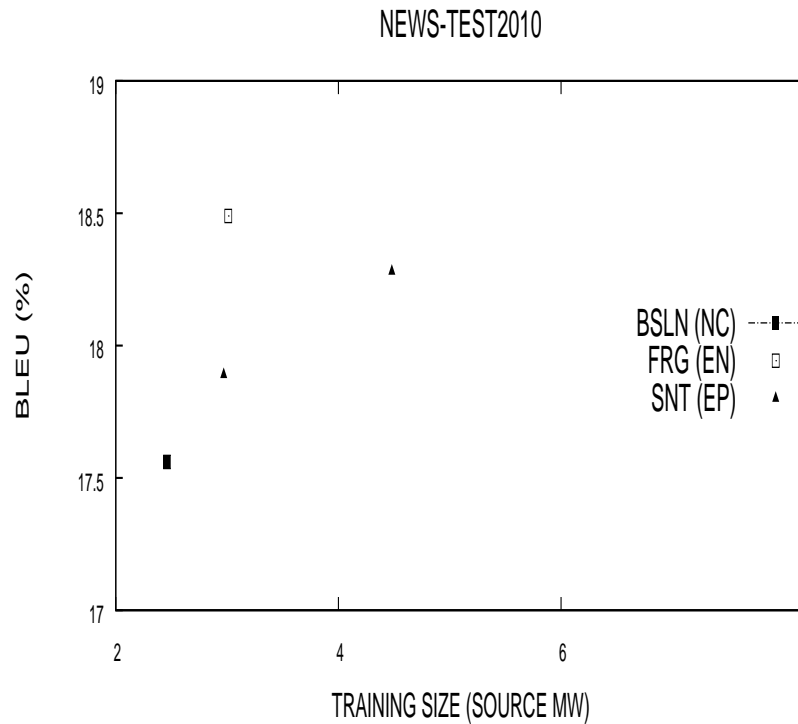


baseline		training data		NEWS TEST 2010
running words (source)	type	running words (source)	type	
2.5M	NC	-	-	17.56
2.5M	NC	0.5M	FRG(EN)	18.49

ACL WMT 2010 translation task: German news translated into English

NC (=)      EN ( $\approx$  in-domain)      EP: EuroParl (= out-of-domain)

SNT is a sentence-based filtering methods from the literature



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

ACL WMT 2010 translation task: German news translated into English

NC (=)      EN ( $\approx$  in-domain)      EP: EuroParl (= out-of-domain)

SNT is a sentence-based filtering methods from the literature

- 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