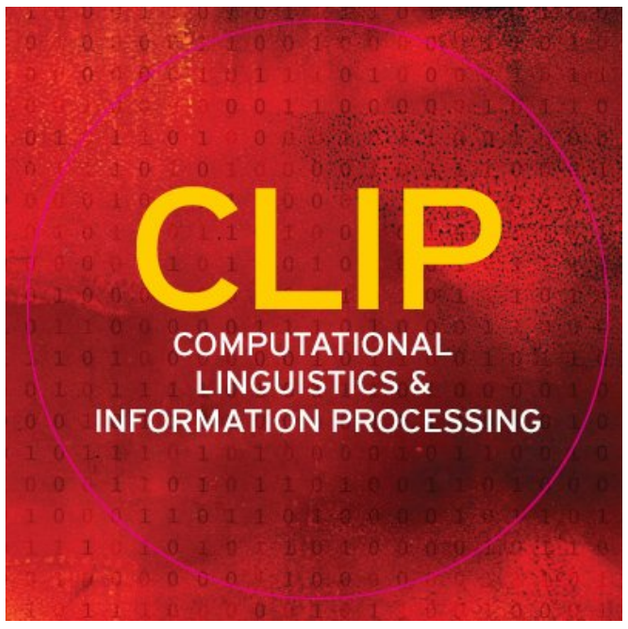
Identifying Semantic Divergences in Parallel Text without Manual Annotations

Yogarshi Vyas, Xing Niu, Marine Carpuat

NAACL 2018 - 06/04/2018



How Parallel is Parallel Text?

- Common Assumption : Parallel sentences are equivalent in meaning

How Parallel is Parallel Text?

- Common Assumption : Parallel sentences are equivalent in meaning

I don’t know what I’m going to do.

J’en sais rien
I do not know.

Someone wanted to cook bratwurst

Vous vouliez des saucisses grillées
you wanted some grilled sausages

Semantic Divergences

Parallel sentences where source and target do not convey the same meaning

Semantic Divergences

Parallel sentences where source and target do not convey the same meaning

- Can impact neural MT more than statistical MT
  - ➡ SMT relatively robust to noise, e.g. sentence mis-alignments (Goutte et al., 2012)
  - ➡ NMT more sensitive to the nature of examples (Chen et al., 2016, Belinkov and Bisk, 2018)

Key Findings : Semantic Divergences ..

- Are common in parallel data
  - ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)

Key Findings : Semantic Divergences ..

- Are common in parallel data

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)

- Can be detected without manual annotations using a deep model of bilingual similarity

  ➡ **80 F1** on a crowdsourced sample

Key Findings : Semantic Divergences ..

- Are common in parallel data

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)

- Can be detected without manual annotations using a deep model of bilingual similarity

  ➡ **80 F1** on a crowdsourced sample

- Have a measurable impact on NMT training

  ➡ Discard most divergent examples yields better BLEU, in less training time

Key Findings : Semantic Divergences ..

- **Are common in parallel data**

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)

- Can be detected without manual annotations using a deep model of bilingual similarity

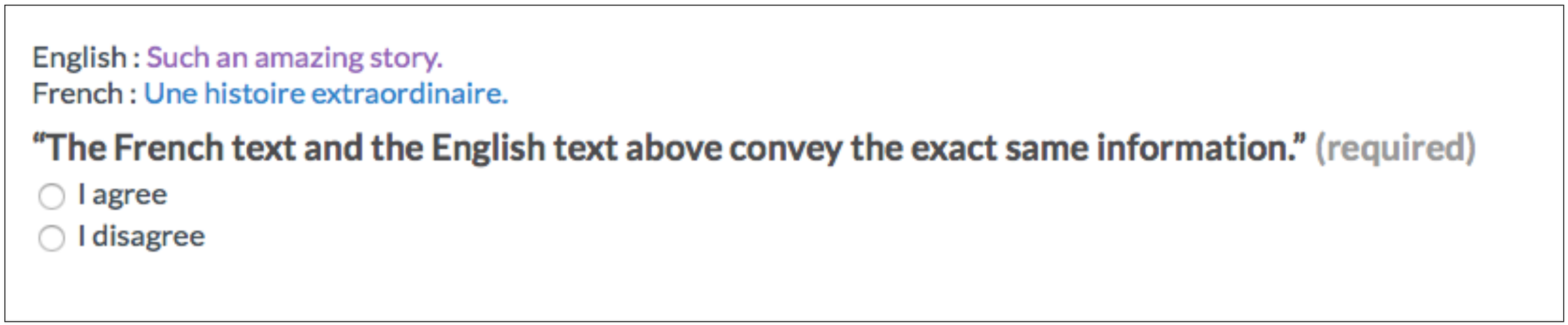
  ➡ **80 F1** on a crowdsourced sample

- Have a measurable impact on NMT training

  ➡ Discard most divergent examples yields better BLEU, in less training time

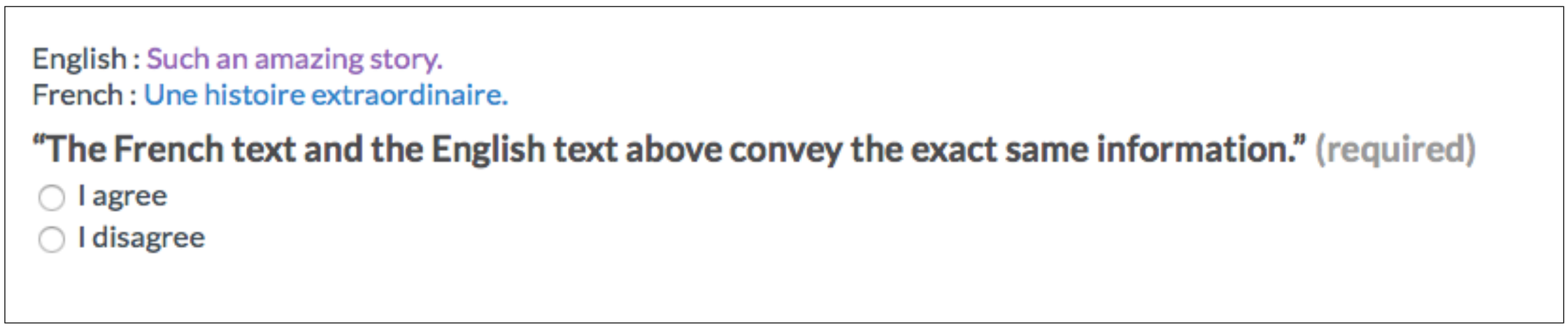
Crowdsourcing reveals divergences are pervasive

- Annotate a sample of 300 sentence pairs from two corpora
  - ➡ Common Crawl and Open Subtitles
  - ➡ 5 bilingual annotators per example

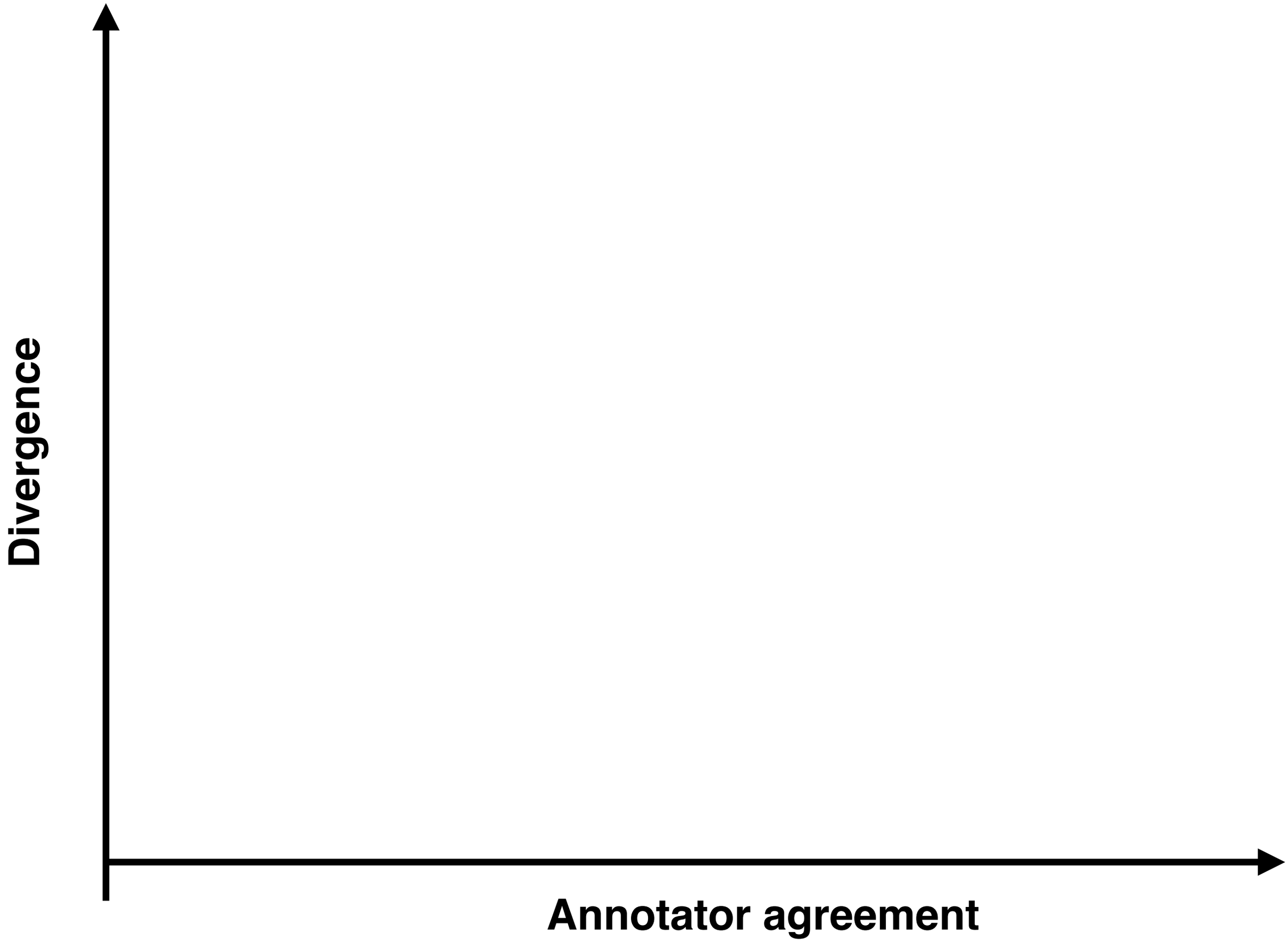


Crowdsourcing reveals divergences are pervasive

- Annotate a sample of 300 sentence pairs from two corpora
  - ➡ Common Crawl and Open Subtitles
  - ➡ 5 bilingual annotators per example



- Majority vote indicates **~40%** pairs are divergent!
  - ➡ ~0.5 Fleiss' Kappa


Divergence
Annotator agreement

Divergence

rabbit? if i told you it was a chicken, you wouldn't know the difference.

vous croirez manger du poulet.

you think you are eating chicken

The epidemic took my wife, my stepson

l’épidémie a touché ma femme, mon beau-fils

the epidemic touched my wife, my stepson

Annotator agreement

Divergence

what does it mean when food is ``low in ash'' or ``low in magnesium''?

quels sont les avantages d'une nourriture ``réduite en cendres'' et ``faible en magnésium'' ?

what are the advantages of a food ``low in ash'' or ``low in magnesium''?

rabbit? if i told you it was a chicken, you wouldn't know the difference.

vous croirez manger du poulet.

you think you are eating chicken

she was a kind person, then, was she?

c'etait quelqu'un de gentil, non ?

it was someone nice, no?

The epidemic took my wife, my stepson

l'épidémie a touché ma femme, mon beau-fils

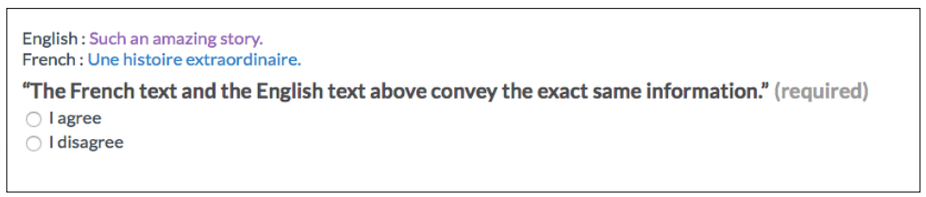
the epidemic touched my wife, my stepson

Annotator agreement

Key Findings : Semantic Divergences ..

- **Are common in parallel data**

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)



- Can be detected without manual annotations using a deep model of bilingual similarity

  ➡ **80 F1** on a crowdsourced sample

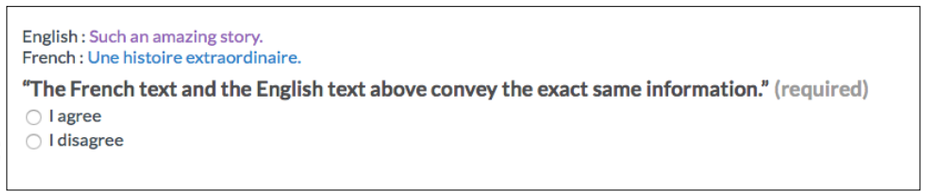
- Have a measurable impact on NMT training

  ➡ Discard most divergent examples yields better BLEU, in less training time

Key Findings : Semantic Divergences ..

- Are common in parallel data

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)



- **Can be detected without manual annotations using a deep model of bilingual similarity**

  ➡ **80 F1** on a crowdsourced sample

- Have a measurable impact on NMT training

  ➡ Discard most divergent examples yields better BLEU, in less training time

Detecting Semantic Divergences

- Identify **similarity** in meaning between parallel sentence pairs

Detecting Semantic Divergences

- Identify **similarity** in meaning between parallel sentence pairs

- Other related tasks :

  ➡ Cross-lingual semantic textual similarity

  ➡ Quality estimation

Detecting Semantic Divergences

- Identify **similarity** in meaning between parallel sentence pairs

- Other related tasks :

  ➡ Cross-lingual semantic textual similarity

  ➡ Quality estimation

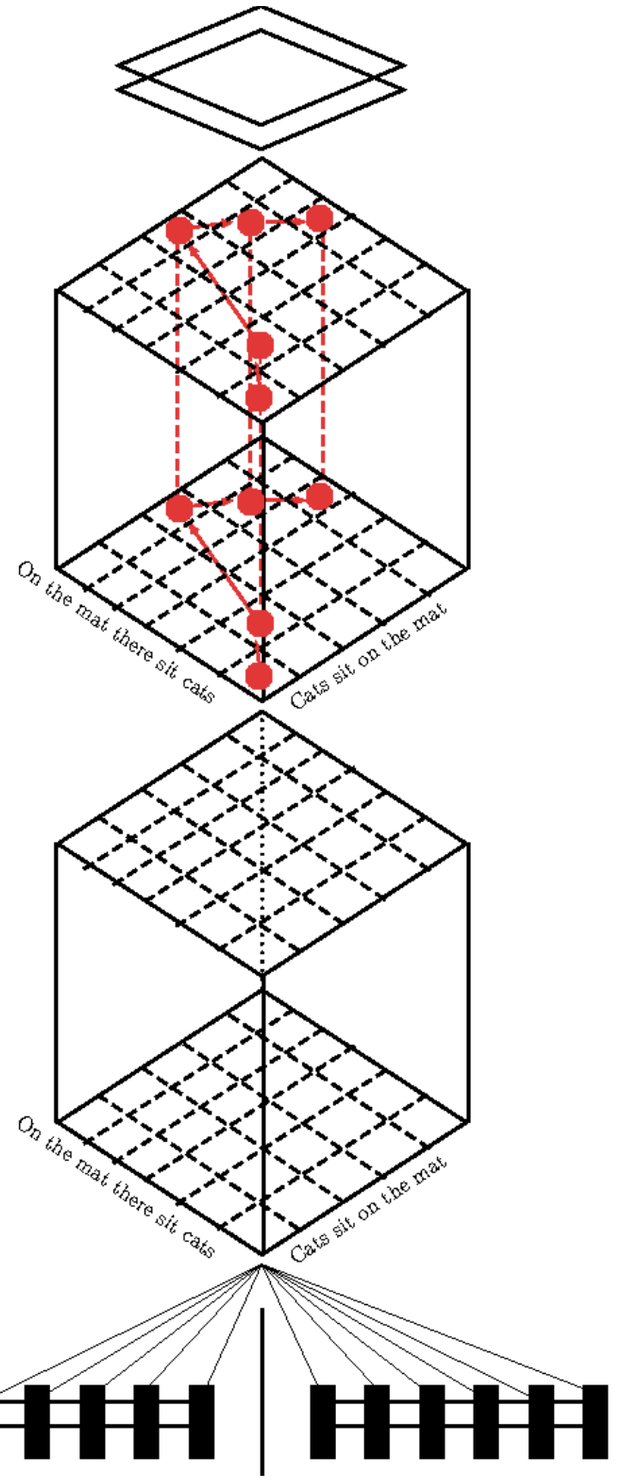
- **Desiderata :** We want a model that

  ➡ Can be applied to any language pair

  ➡ Performs strongly on benchmark tasks

Deep Model of Bilingual Semantic Similarity

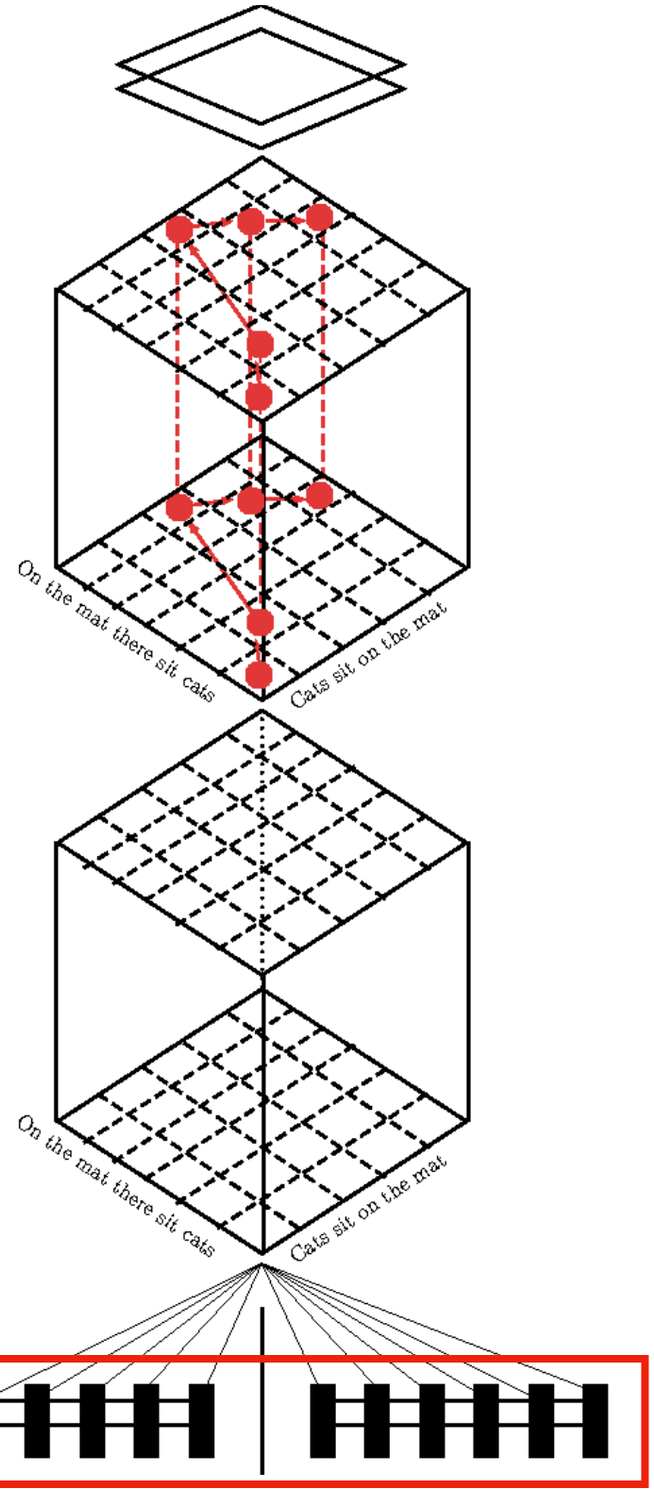
- **Model** : Very Deep Pairwise Word Interaction model
  (He and Lin, 2016)



Deep Model of Bilingual Semantic Similarity

- **Model** : Very Deep Pairwise Word Interaction model
  (He and Lin, 2016)

  ➡ RNNs for contextual word meaning

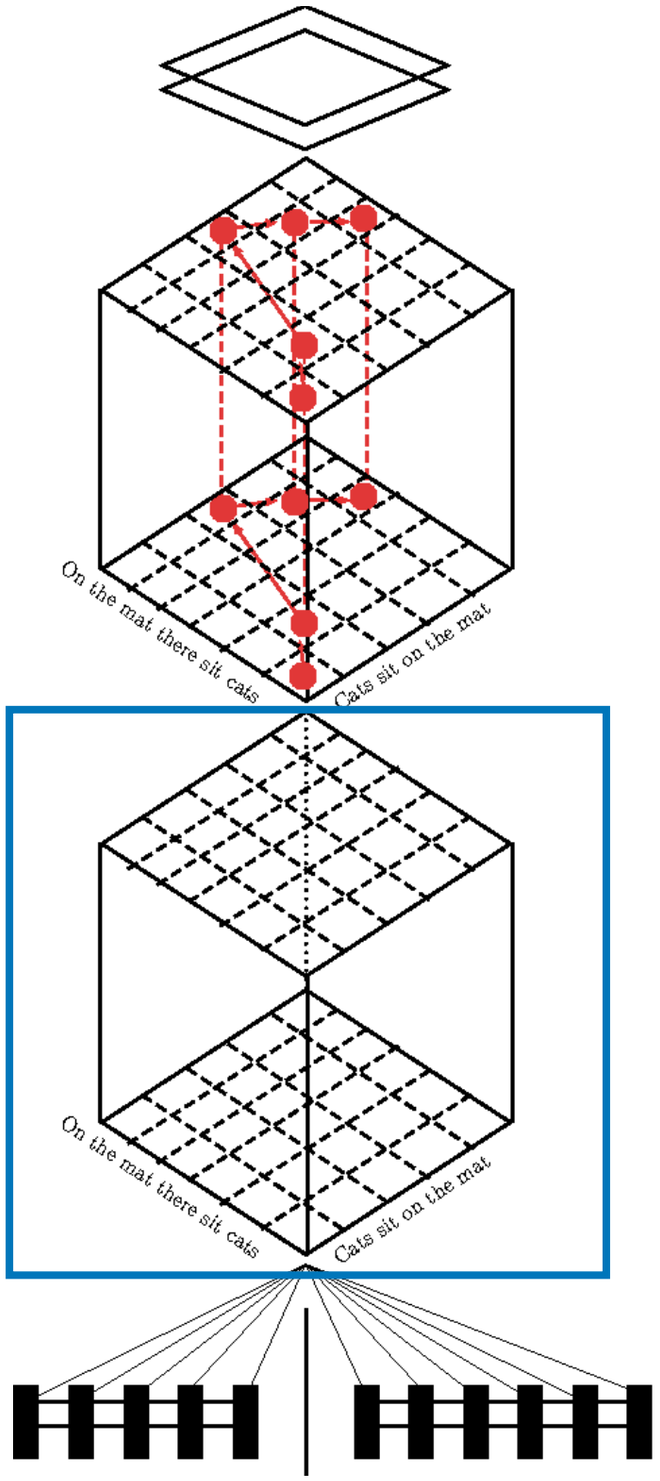


Deep Model of Bilingual Semantic Similarity

- **Model** : Very Deep Pairwise Word Interaction model
  (He and Lin, 2016)

  ➡ RNNs for contextual word meaning

  ➡ Pairwise word similarities



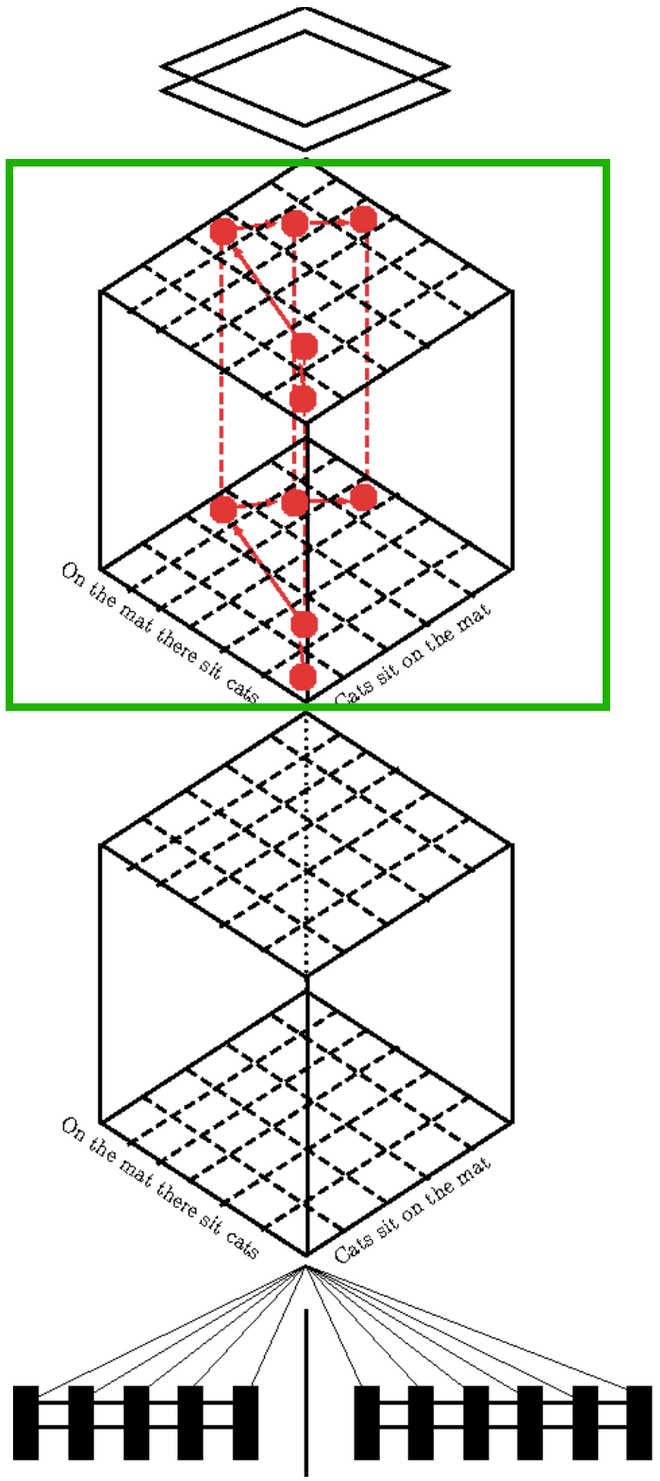
Deep Model of Bilingual Semantic Similarity

- **Model** : Very Deep Pairwise Word Interaction model
  (He and Lin, 2016)

  ➡ RNNs for contextual word meaning

  ➡ Pairwise word similarities

  ➡ Up-weigh important word pairs



Deep Model of Bilingual Semantic Similarity

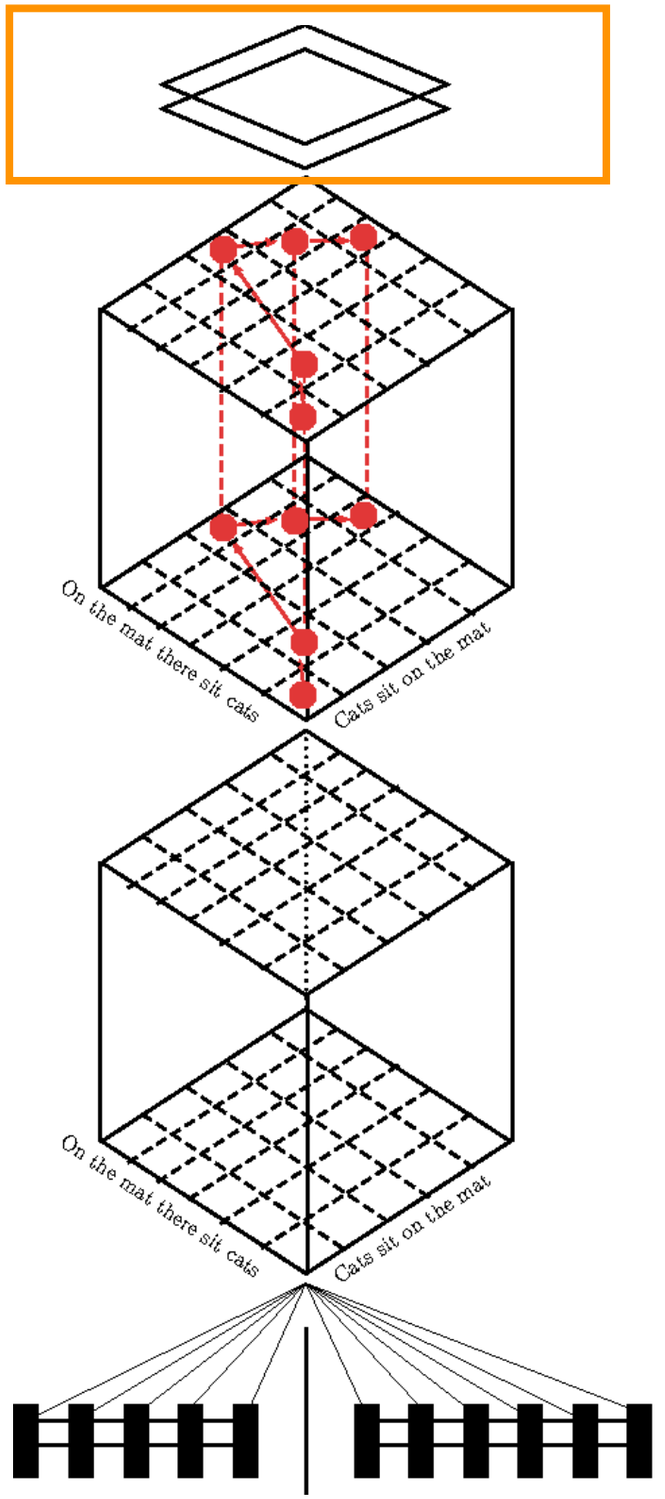
- **Model** : Very Deep Pairwise Word Interaction model
  (He and Lin, 2016)

  ➡ RNNs for contextual word meaning

  ➡ Pairwise word similarities

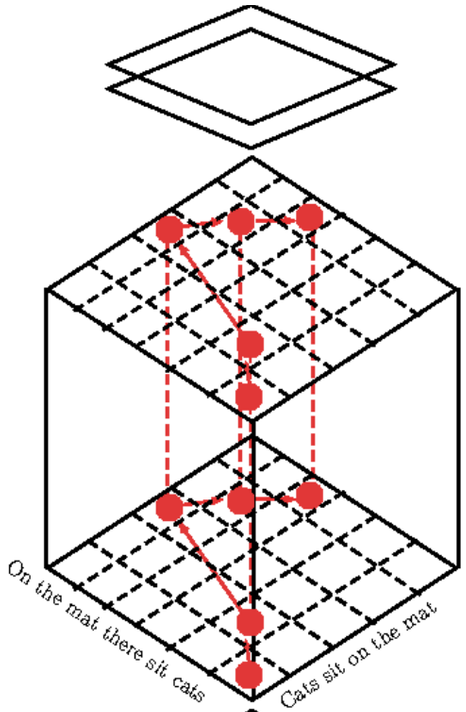
  ➡ Up-weigh important word pairs

  ➡ CNN for feature extraction and classification

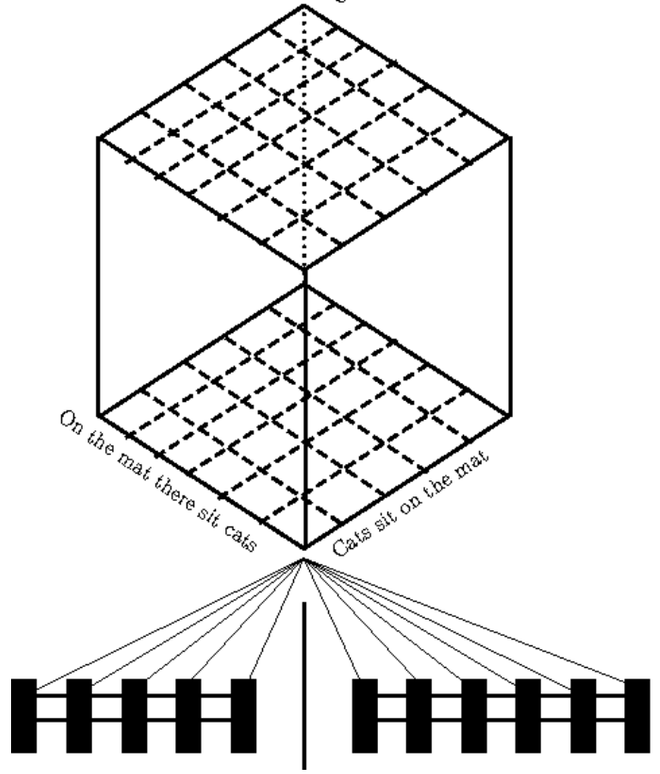


Deep Model of Bilingual Semantic Similarity

- **Model** : Very Deep Pairwise Word Interaction model
  (He and Lin, 2016)
  - ➡ RNNs for contextual word meaning
  - ➡ Pairwise word similarities
  - ➡ Up-weigh important word pairs
  - ➡ CNN for feature extraction and classification

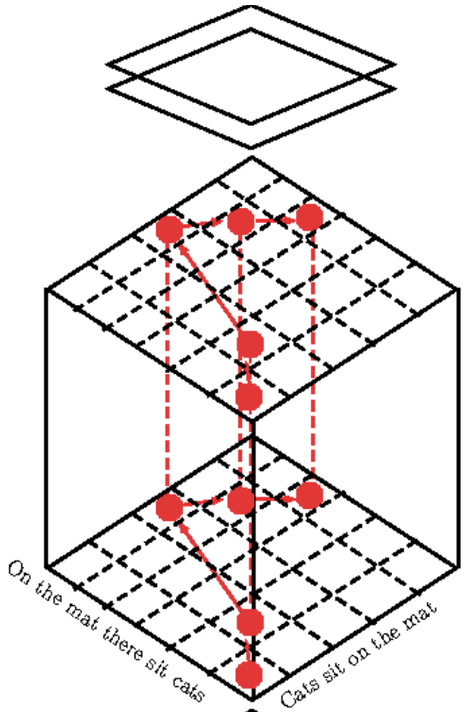


- **Challenges**
  - ➡ How to use model for bilingual sentences?
  - ➡ What data can we use to train the model?

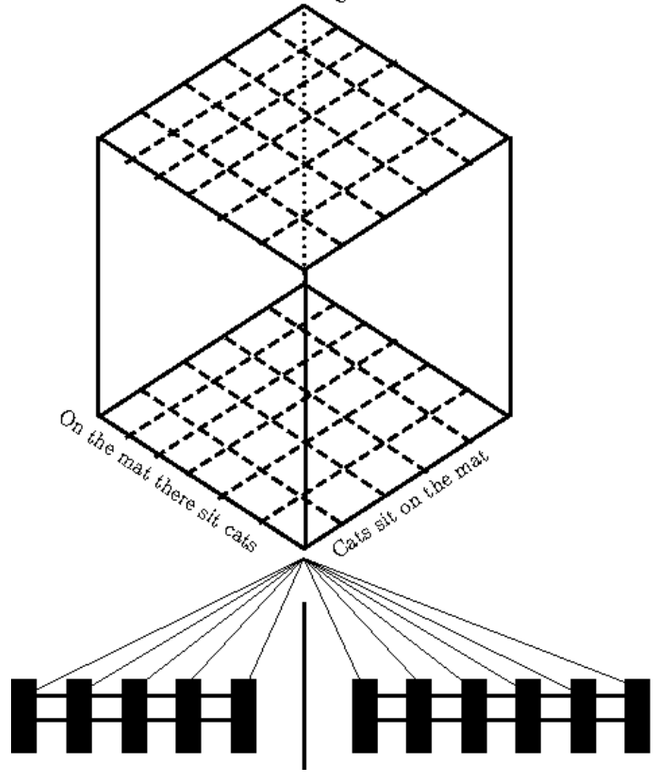


Deep Model of Bilingual Semantic Similarity

- **Model** : Very Deep Pairwise Word Interaction model
  (He and Lin, 2016)
  - ➡ RNNs for contextual word meaning
  - ➡ Pairwise word similarities
  - ➡ Up-weigh important word pairs
  - ➡ CNN for feature extraction and classification



- **Challenges**
  - ➡ How to use model for bilingual sentences?
    - ▸ Initialize with bilingual embeddings
  - ➡ What data can we use to train the model?



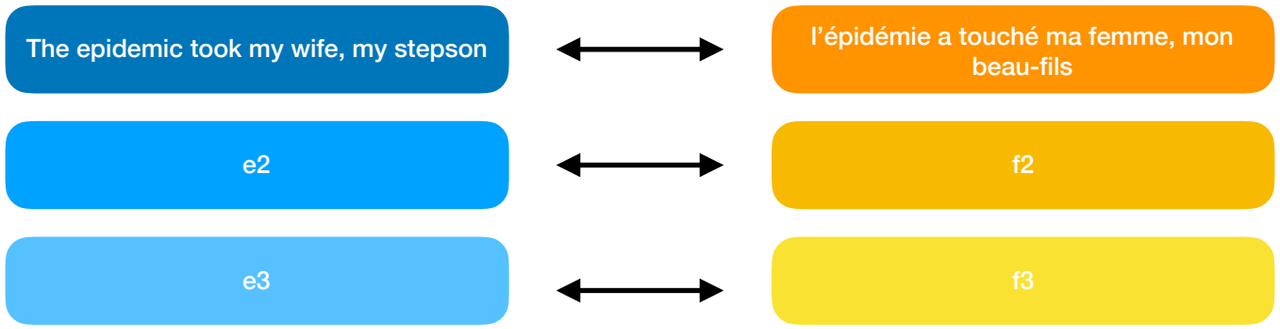
Noisy Synthetic Supervision

+

-

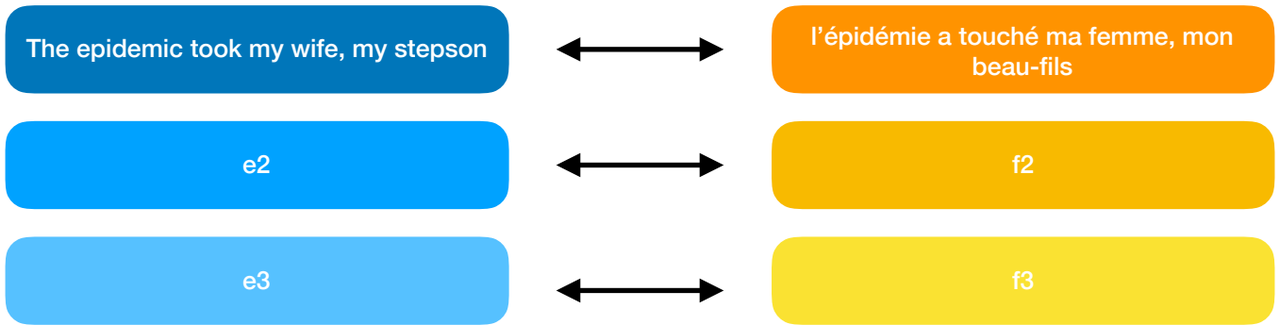
(Munteanu and Marcu 2005)

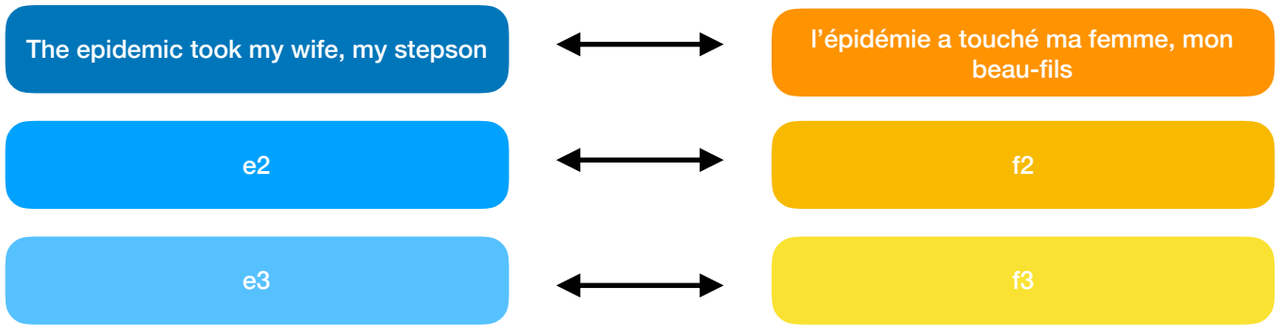
Noisy Synthetic Supervision



(Munteanu and Marcu 2005)

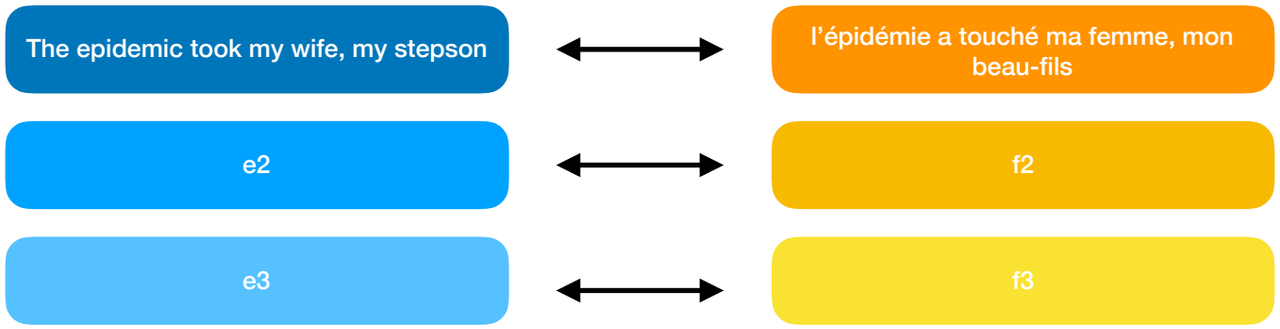
Noisy Synthetic Supervision

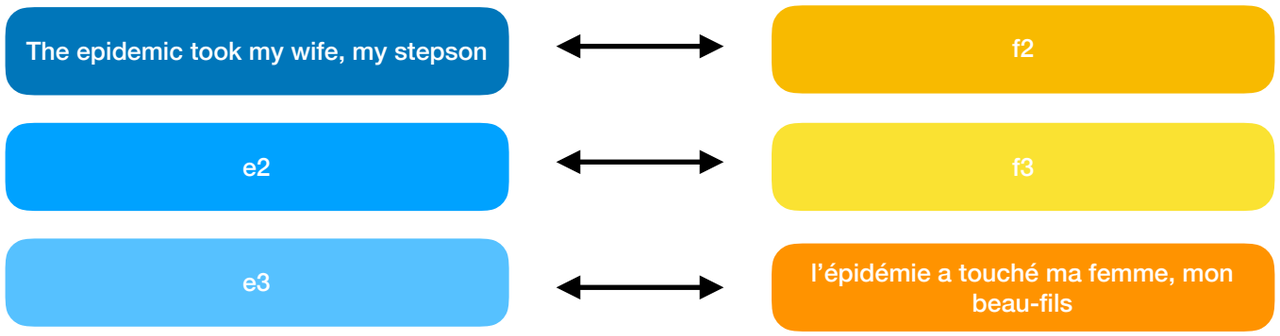




(Munteanu and Marcu 2005)

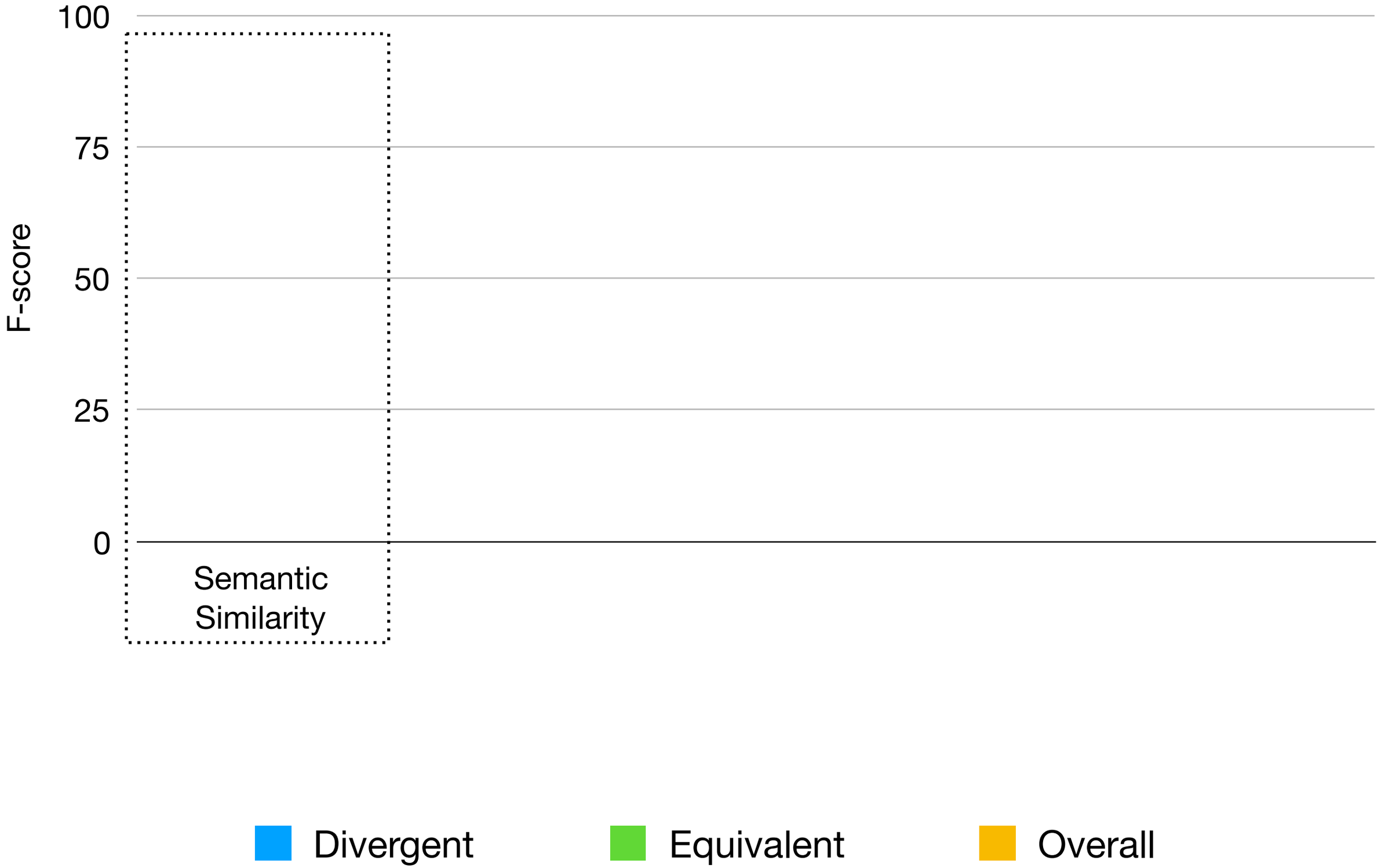
Noisy Synthetic Supervision



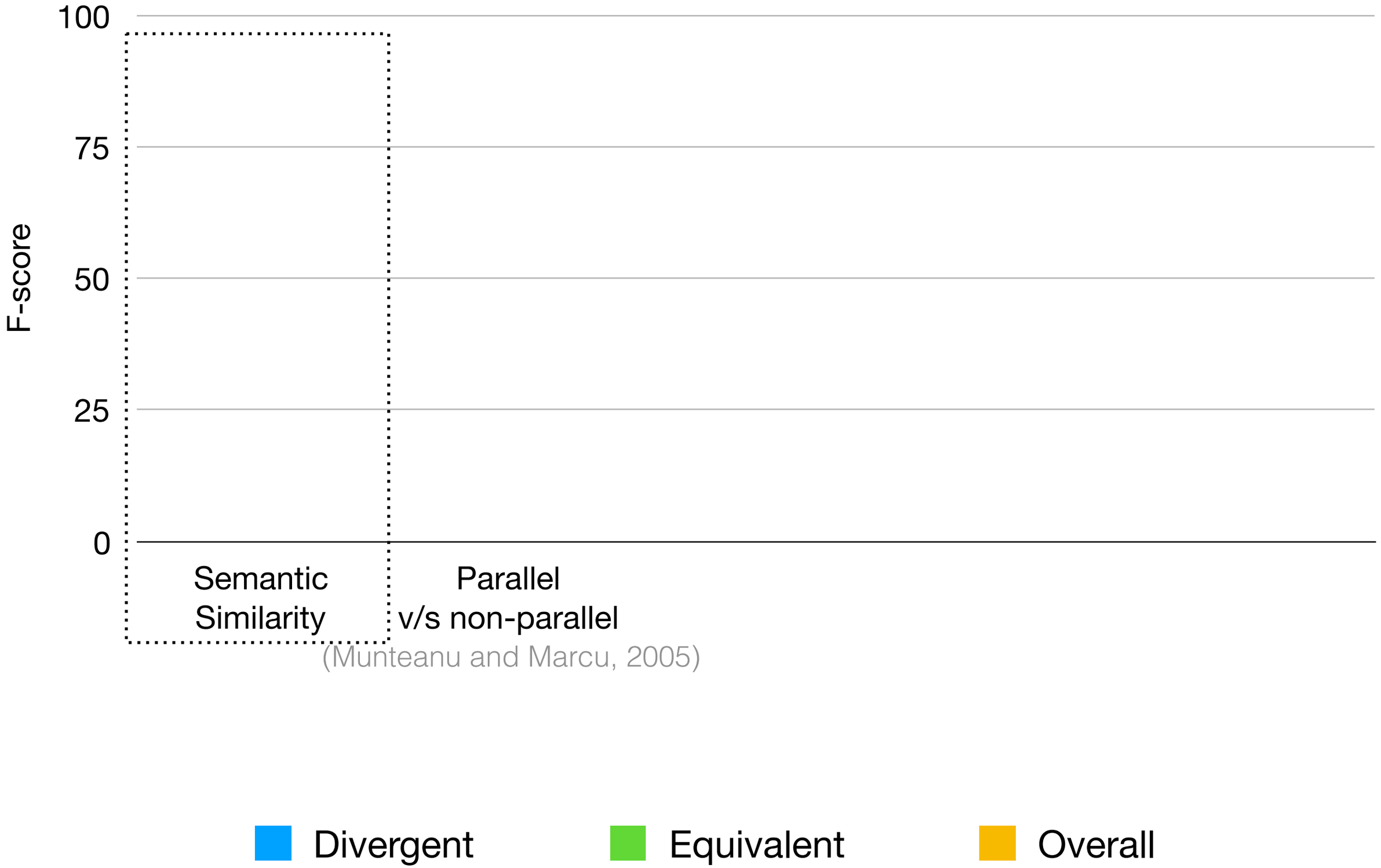


(Munteanu and Marcu 2005)

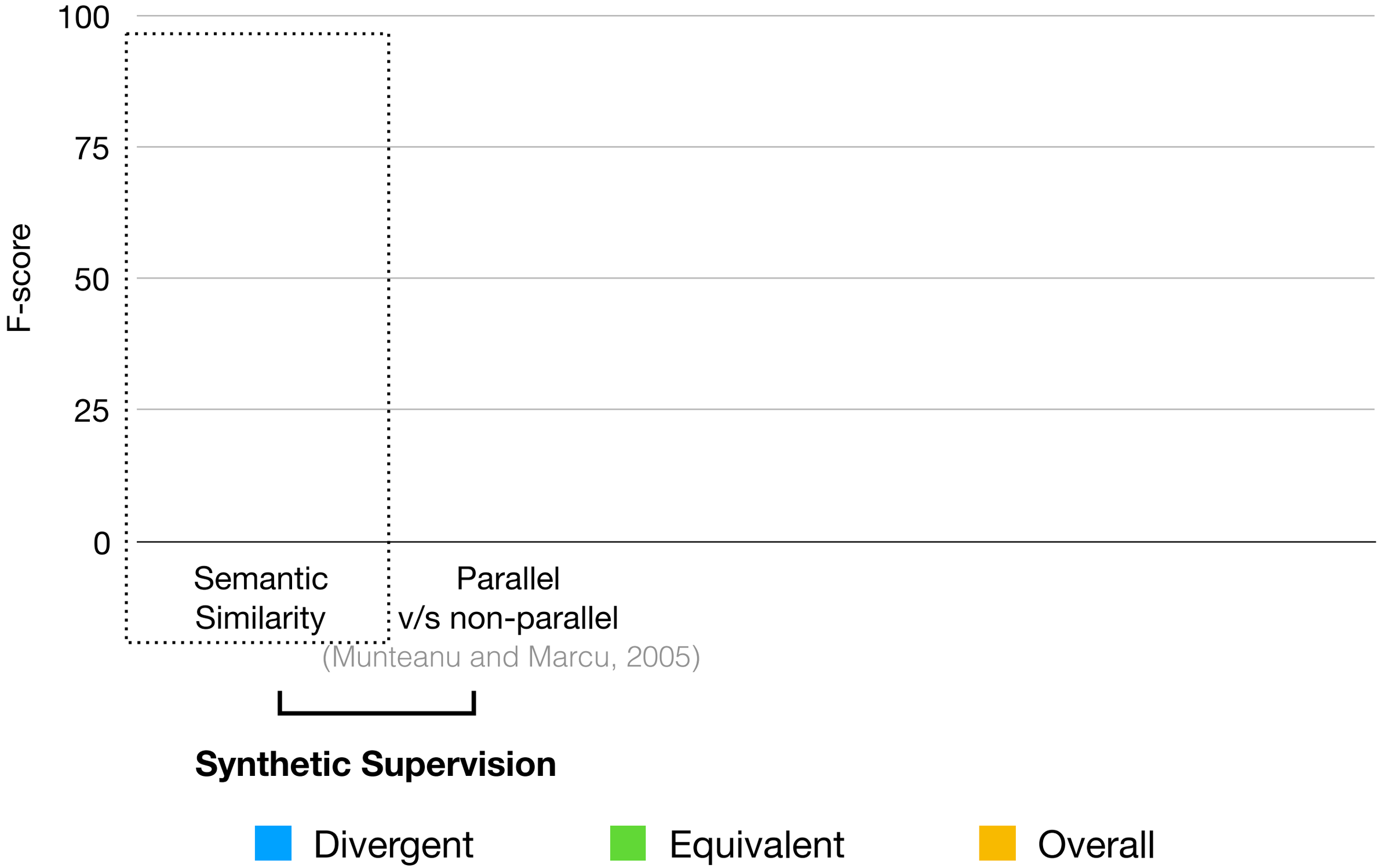
Intrinsic Detection of Divergences (OpenSubtitles)



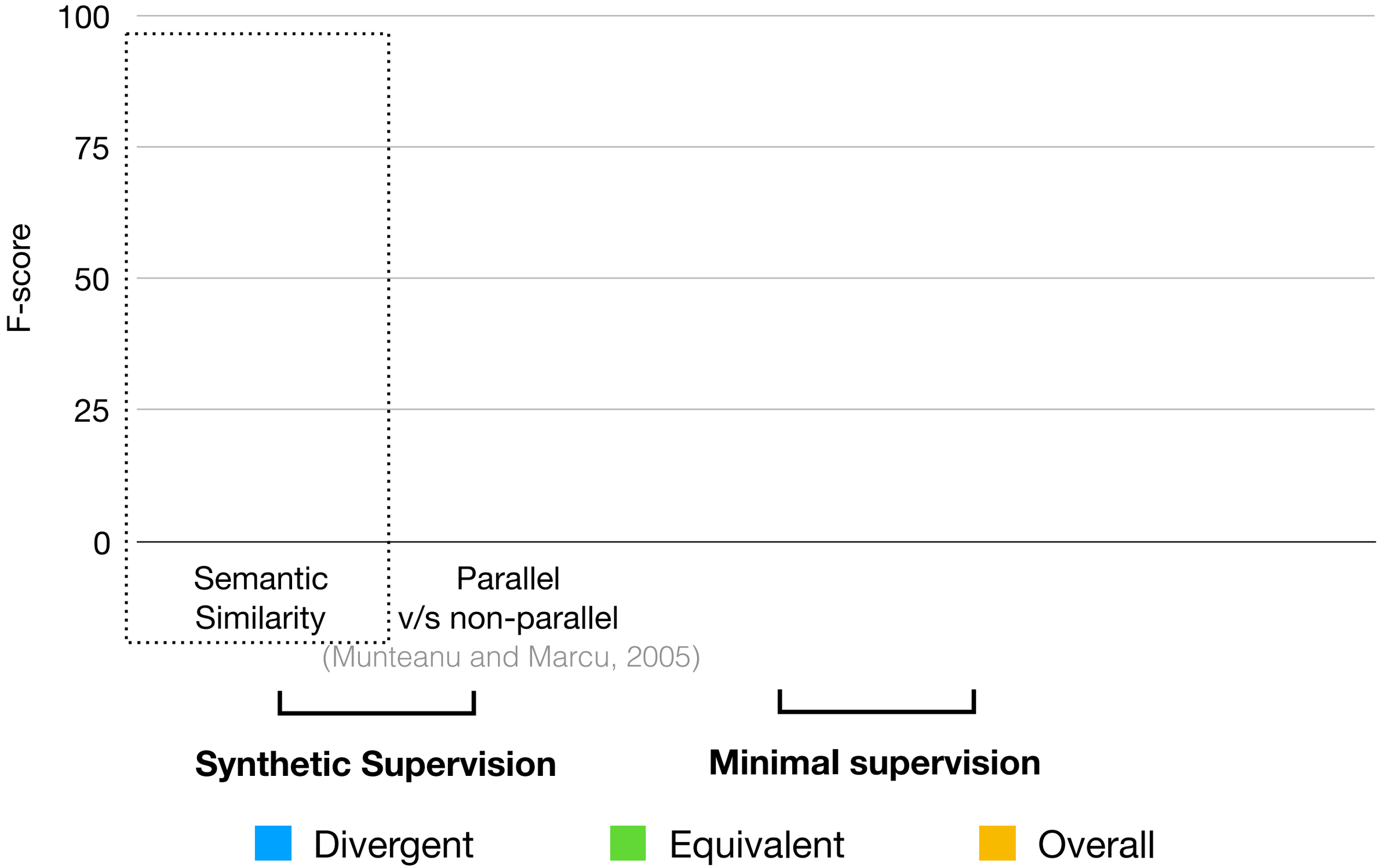
Intrinsic Detection of Divergences (OpenSubtitles)



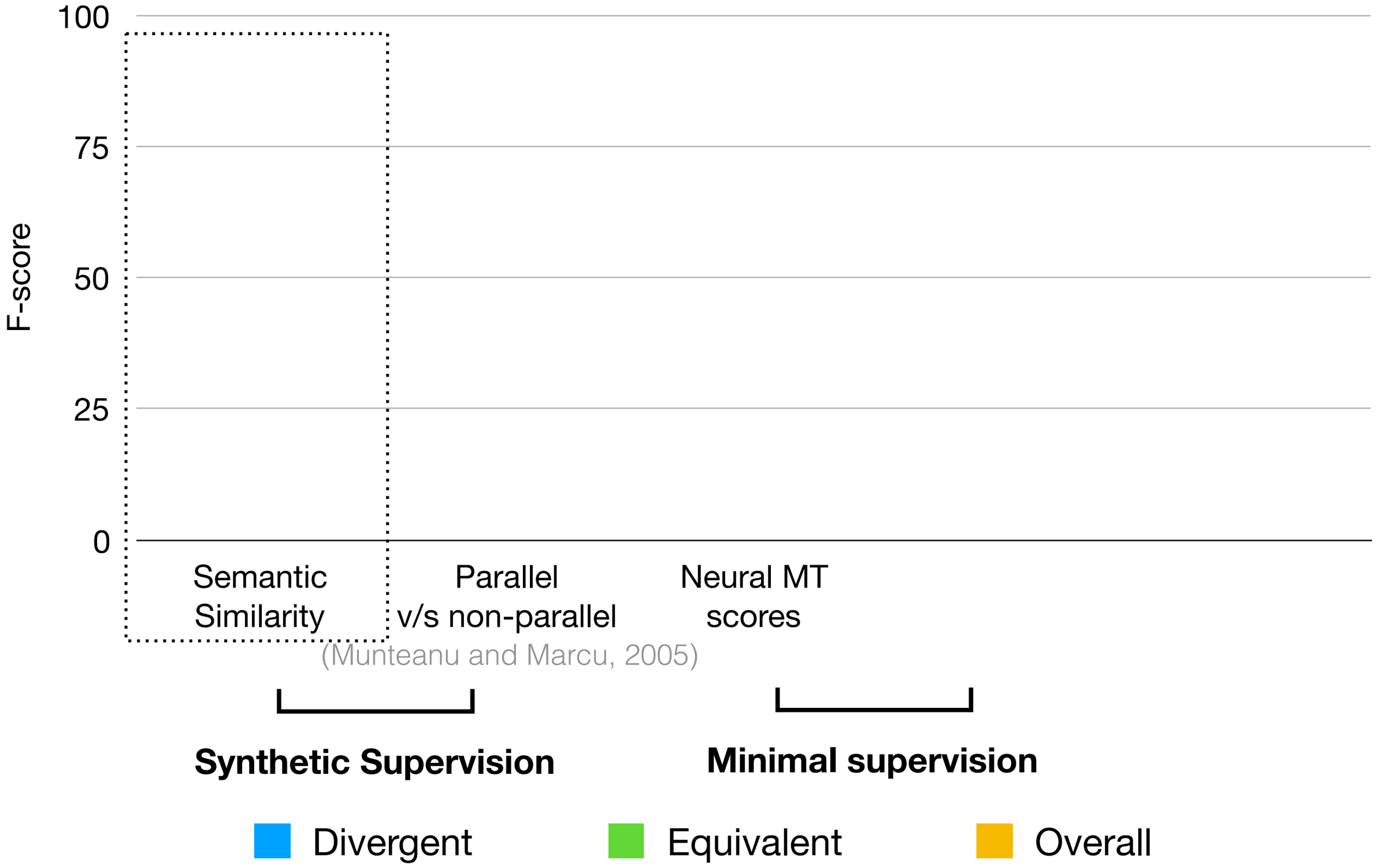
Intrinsic Detection of Divergences (OpenSubtitles)



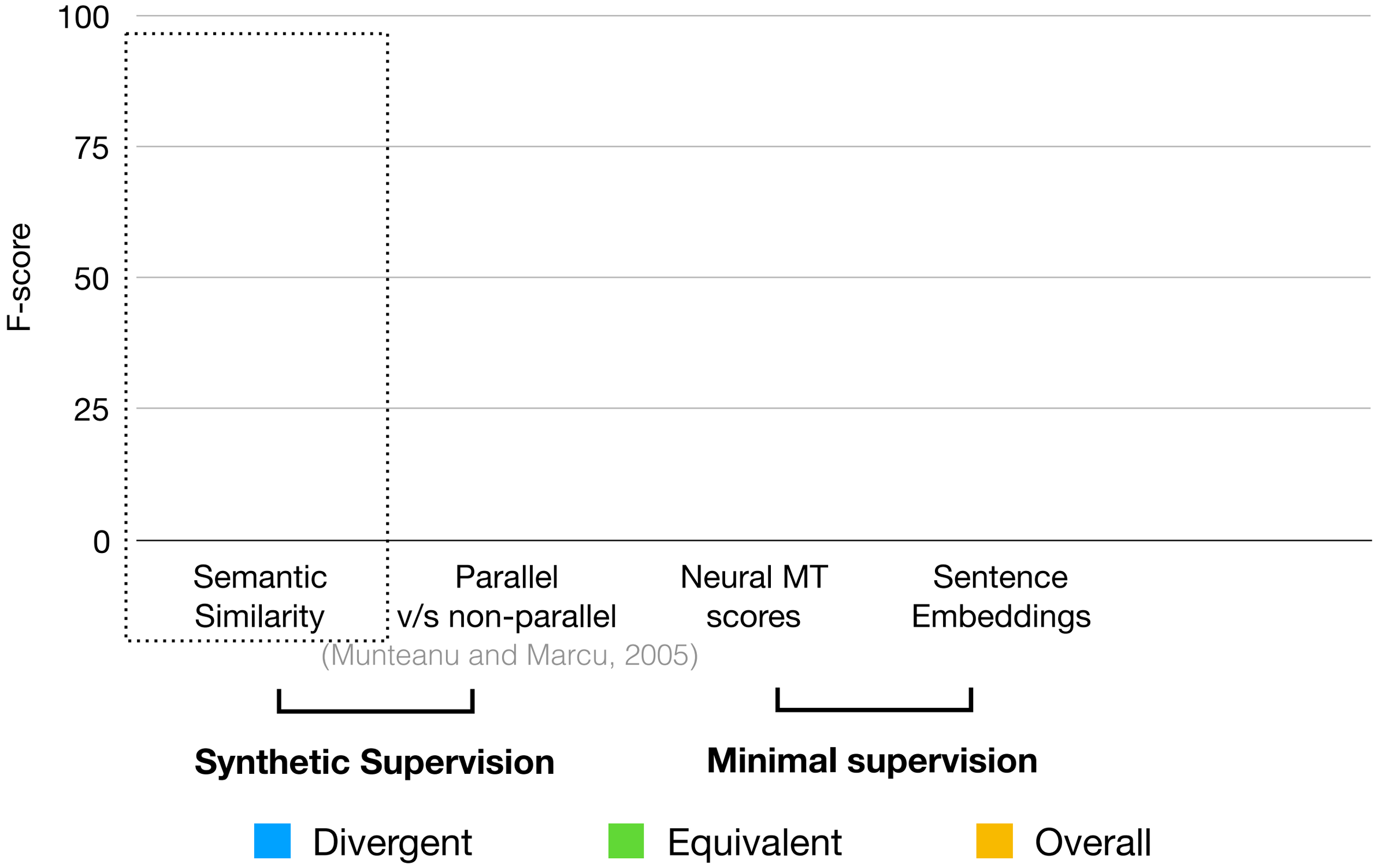
Intrinsic Detection of Divergences (OpenSubtitles)



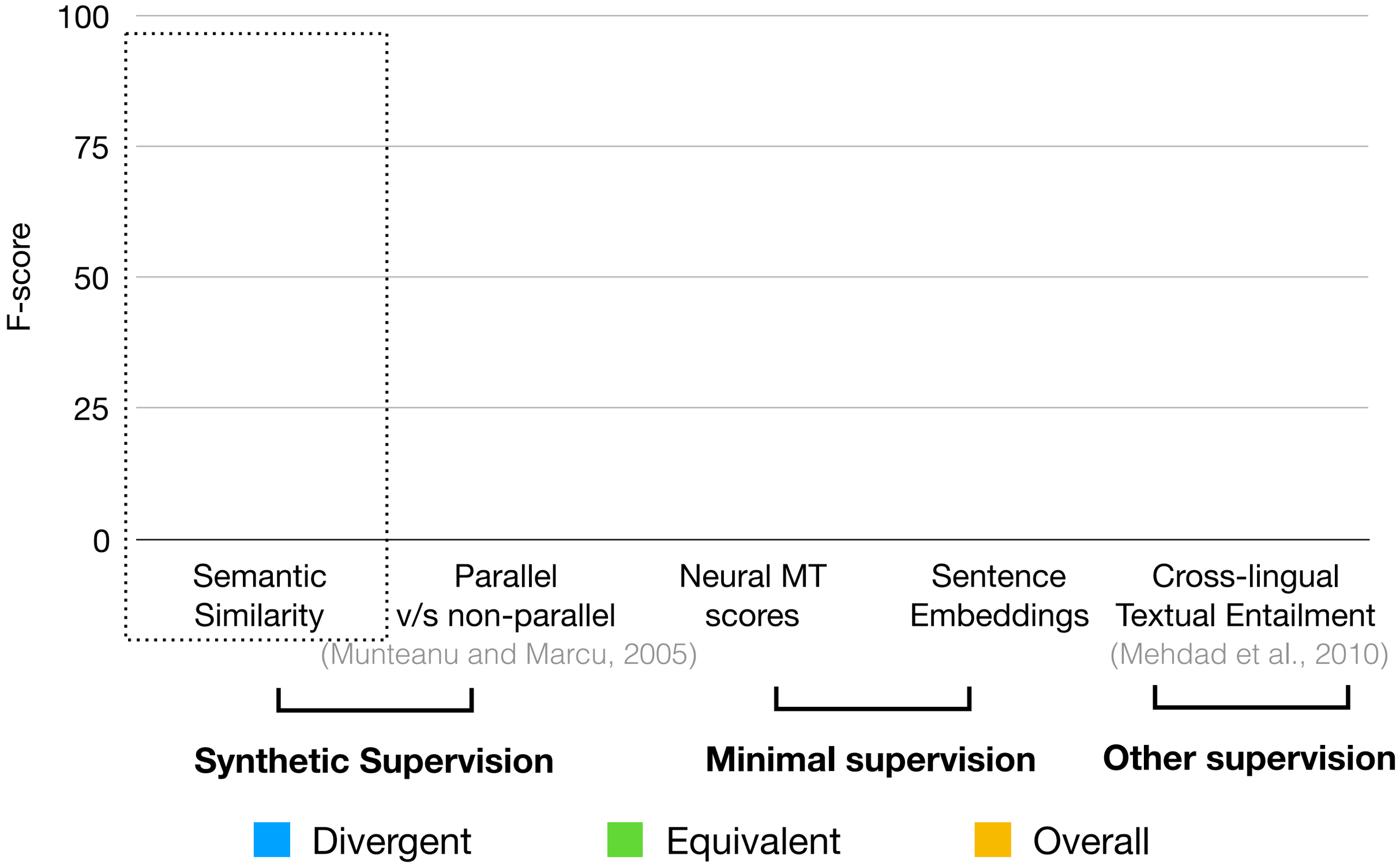
Intrinsic Detection of Divergences (OpenSubtitles)



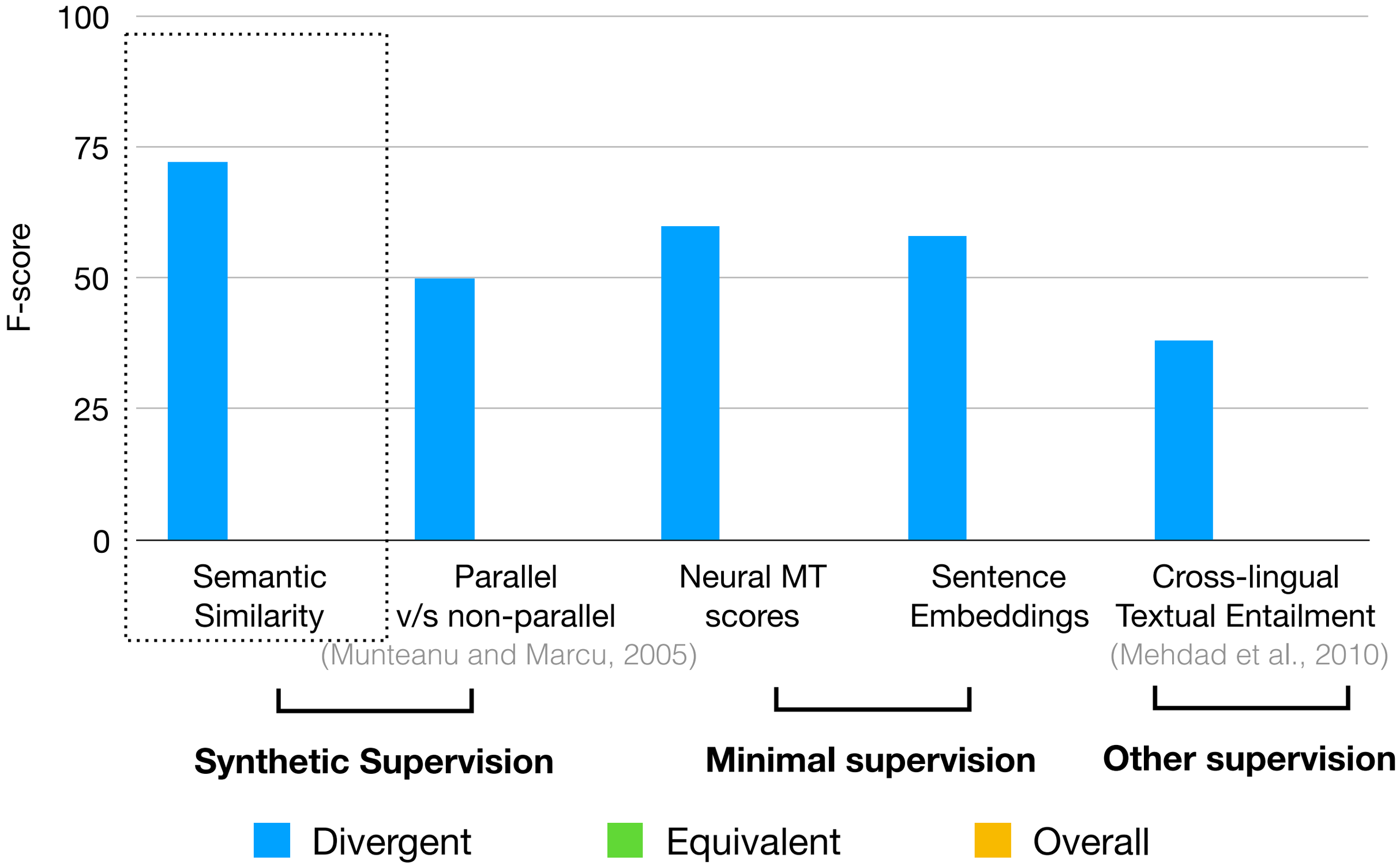
Intrinsic Detection of Divergences (OpenSubtitles)



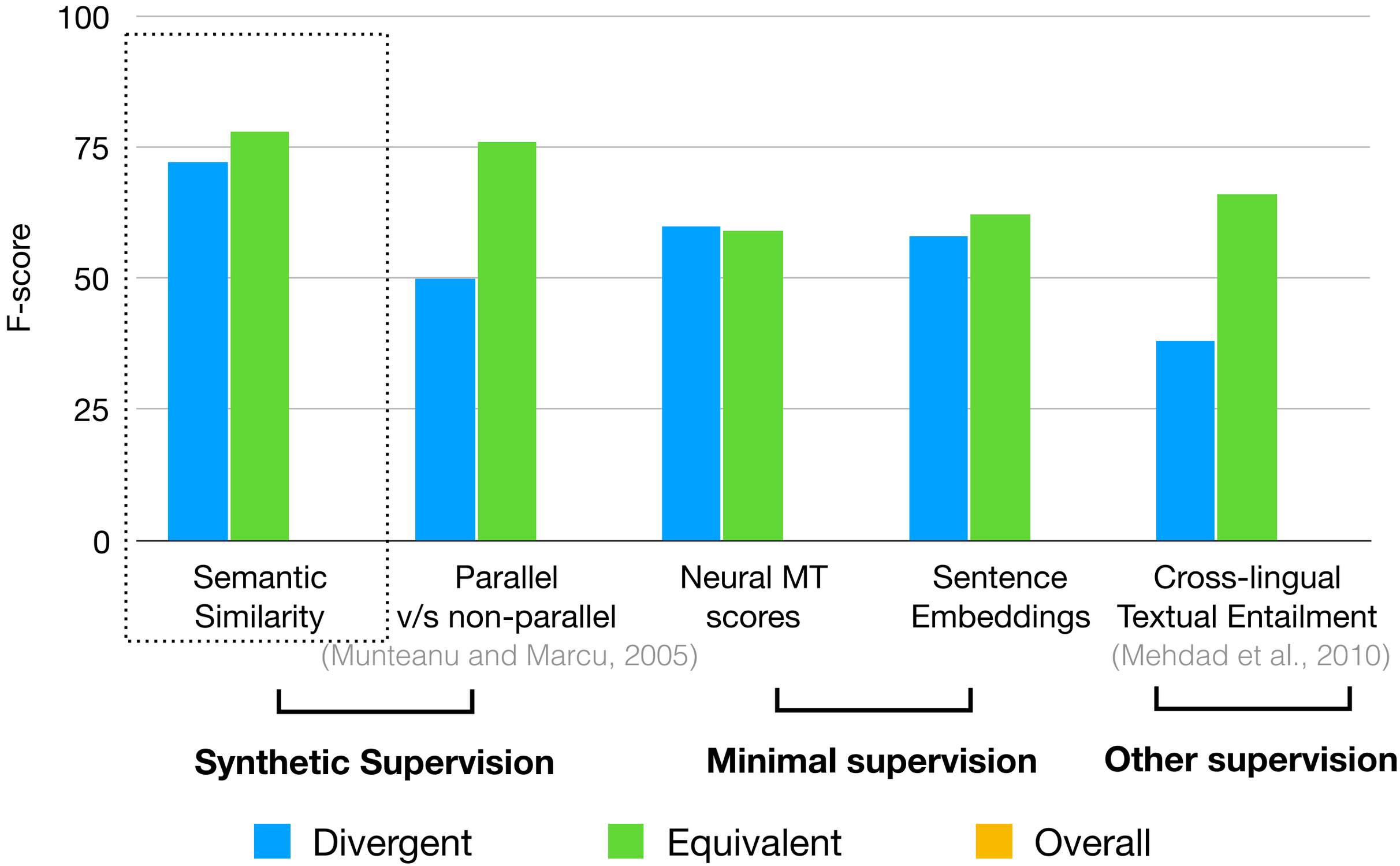
Intrinsic Detection of Divergences (OpenSubtitles)



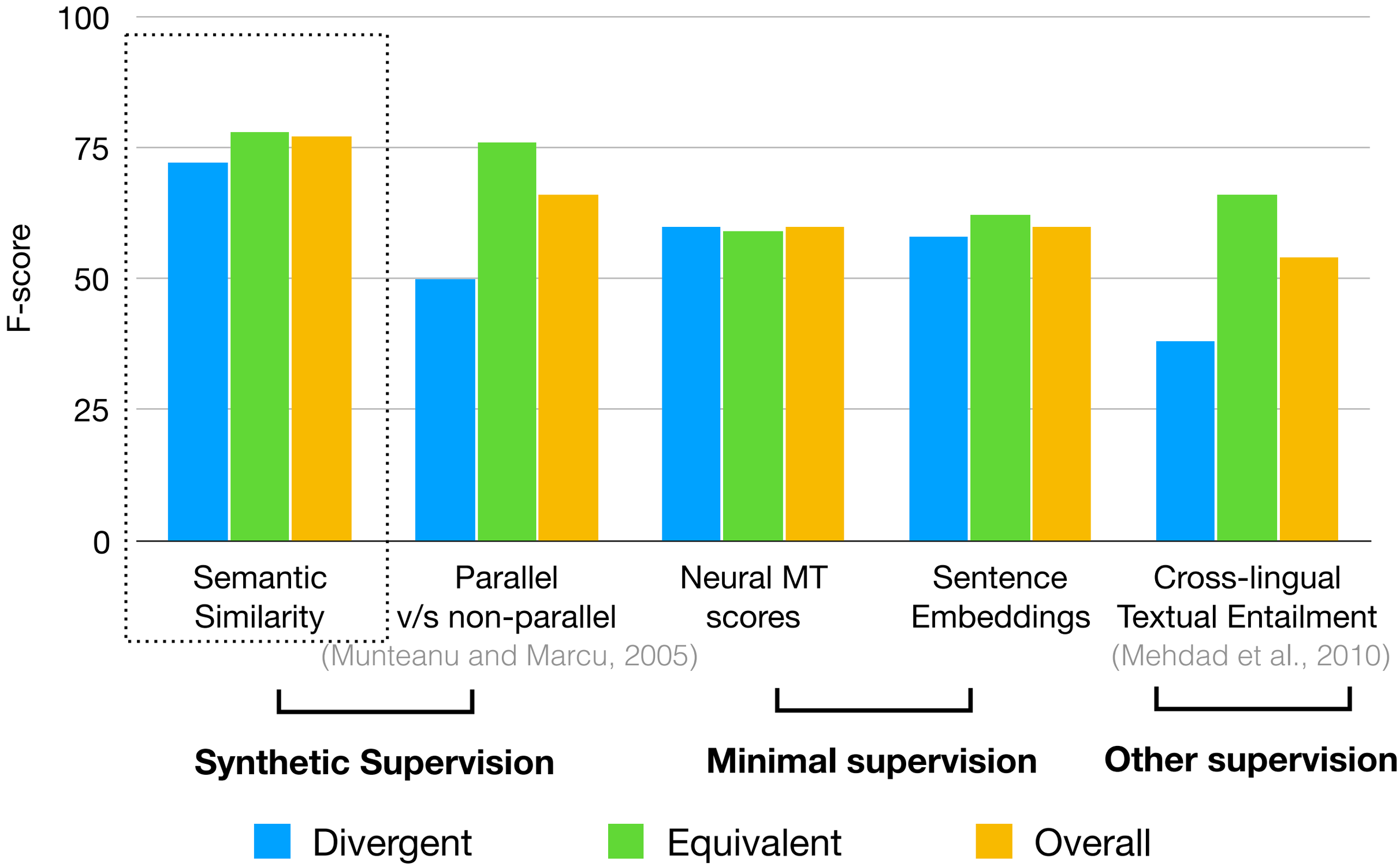
Intrinsic Detection of Divergences (OpenSubtitles)



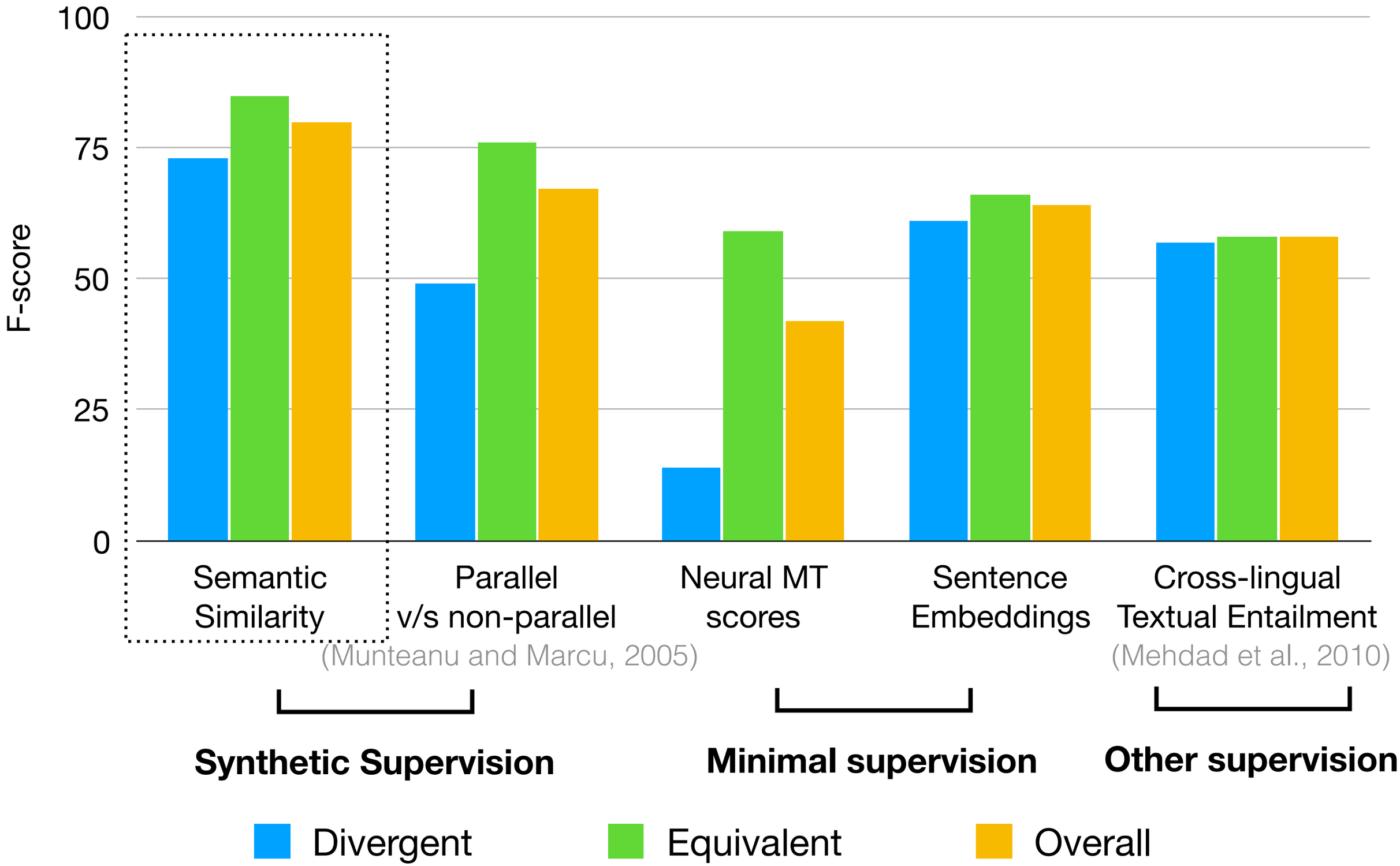
Intrinsic Detection of Divergences (OpenSubtitles)



Intrinsic Detection of Divergences (OpenSubtitles)



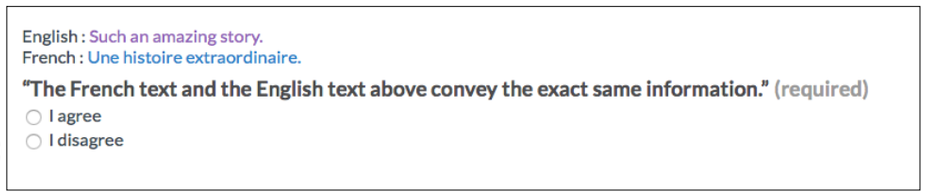
Intrinsic Detection of Divergences (Common Crawl)



Key Findings : Semantic Divergences ..

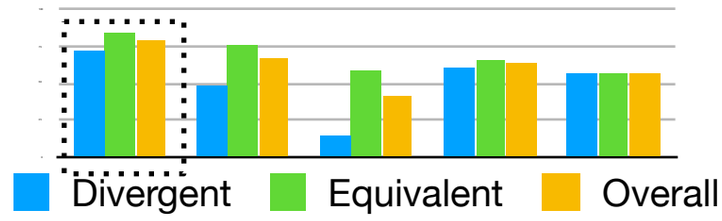
- Are common in parallel data

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)



- **Can be detected without manual annotations using a deep model of bilingual similarity**

  ➡ **80 F1** on a crowdsourced sample



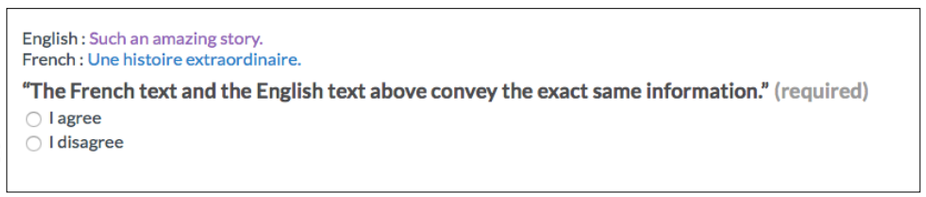
- Have a measurable impact on NMT training

  ➡ Discard most divergent examples yields better BLEU, in less training time

Key Findings : Semantic Divergences ..

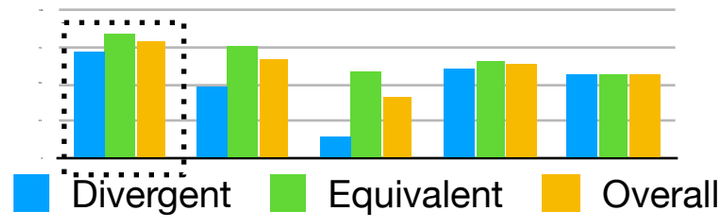
- Are common in parallel data

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)



- Can be detected without manual annotations using a deep model of bilingual similarity

  ➡ **80 F1** on a crowdsourced sample



- **Have a measurable impact on NMT training**

  ➡ Discard most divergent examples yields better BLEU, in less training time

Measuring Impact on NMT by Data Selection

- **Data Selection** : Select data that is labeled least divergent

Measuring Impact on NMT by Data Selection

- **Data Selection** : Select data that is labeled least divergent

- **Translation Task :**

  ➡ English-French

  ➡ Vietnamese-English

Measuring Impact on NMT by Data Selection

- **Data Selection** : Select data that is labeled least divergent

- **Translation Task :**

  ➡ English-French

  - Train : OpenSubtitles (28M —> 14M)
  - Test : Microsoft Spoken Language Translation, TED Talks

  ➡ Vietnamese-English

  - Train : TED Talks (0.12M —> 0.10M)
  - Test : TED Talks

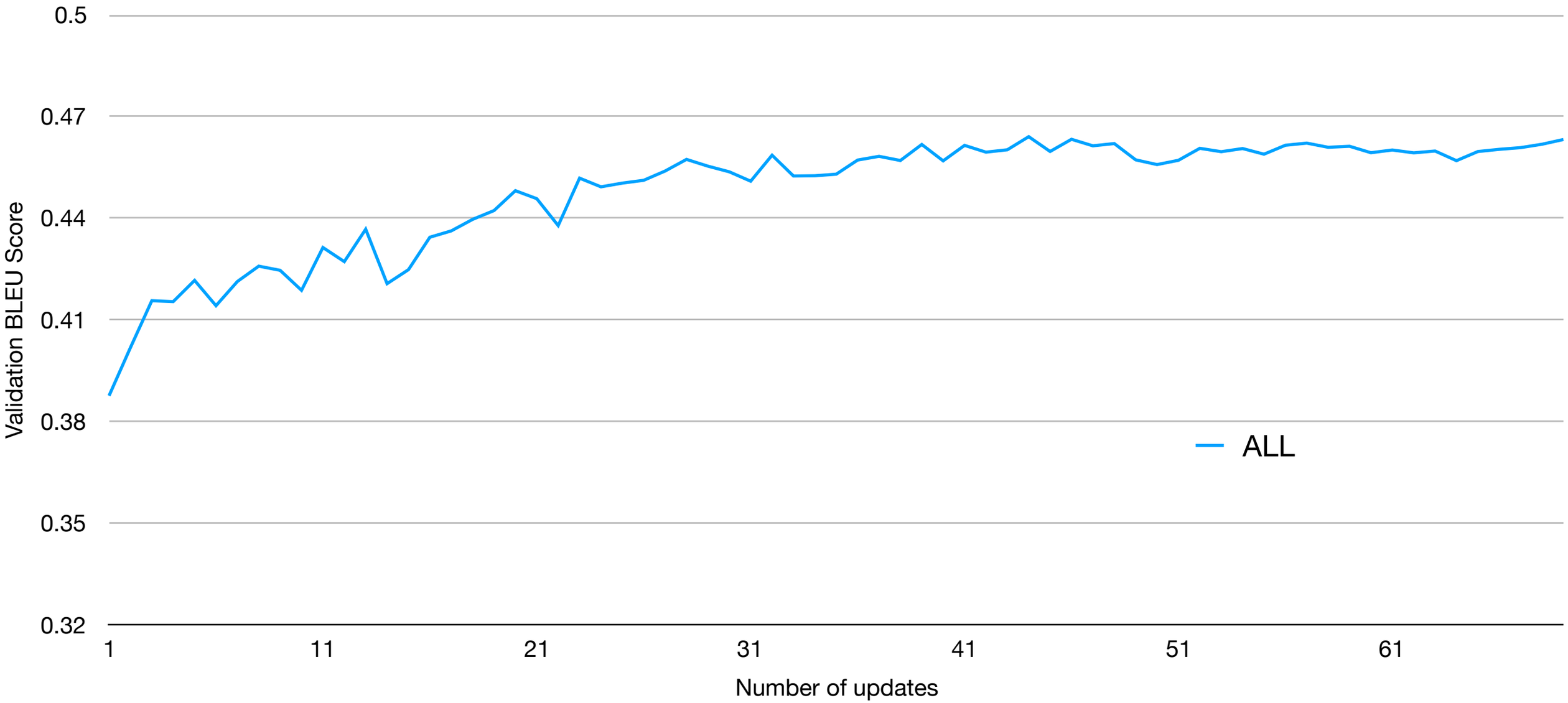
Measuring Impact on NMT by Data Selection

- **Data Selection** : Select data that is labeled least divergent

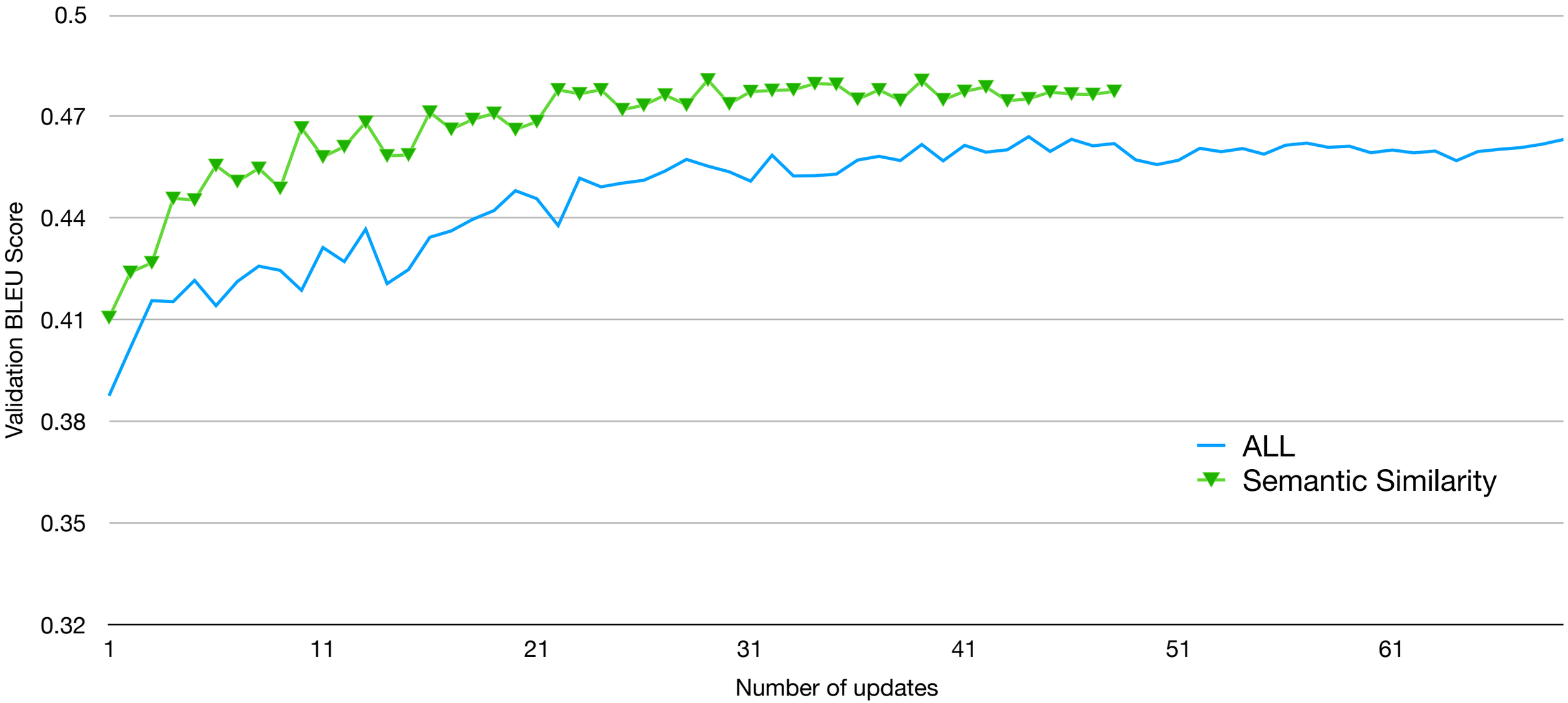
- **Translation Task :**
  - ➡ English-French
    - Train : OpenSubtitles (28M —> 14M)
    - Test : Microsoft Spoken Language Translation, TED Talks
  - ➡ Vietnamese-English
    - Train : TED Talks (0.12M —> 0.10M)
    - Test : TED Talks

- **Contrastive Approaches for data selection**
  - ➡ Parallel v/s Non-parallel
  - ➡ Cross-lingual Textual Entailment

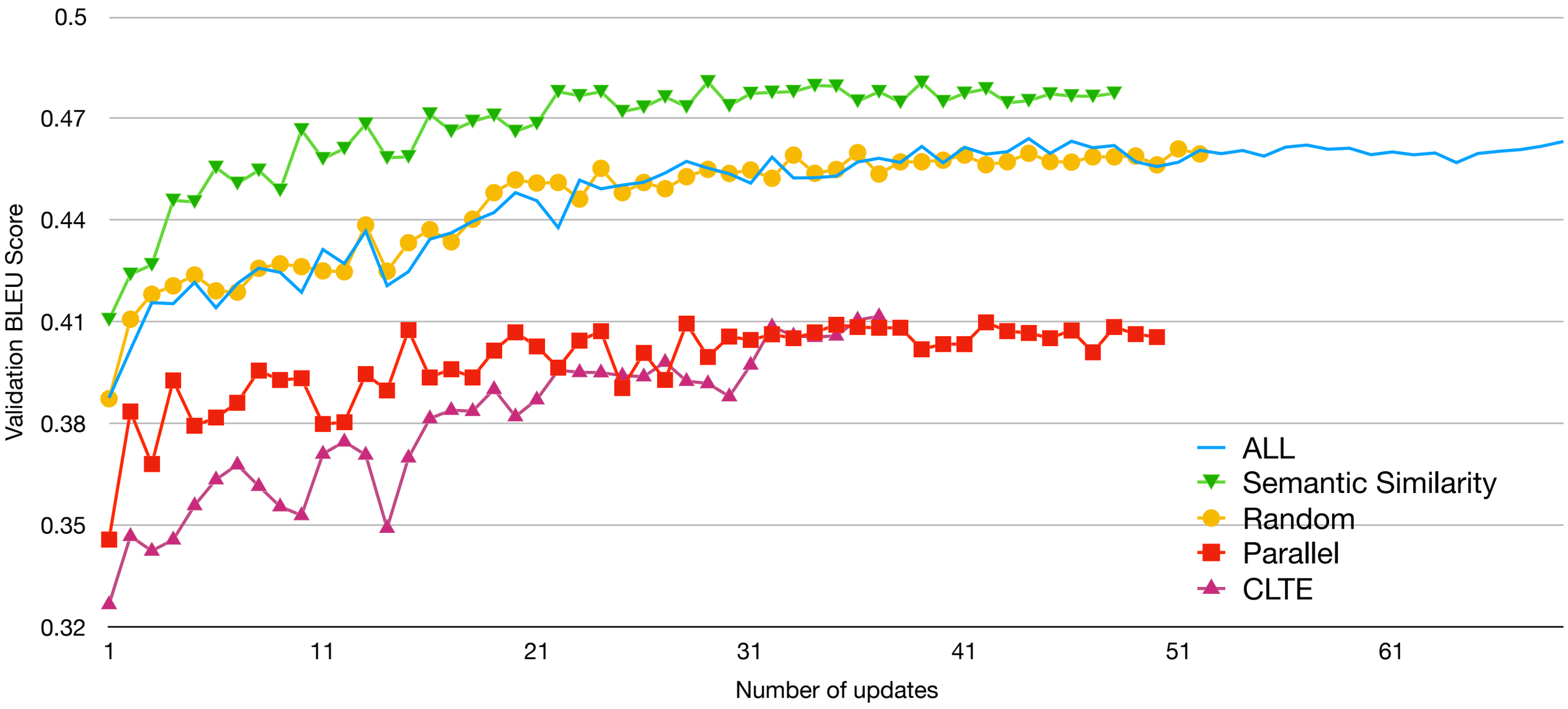
Faster and Better NMT Training



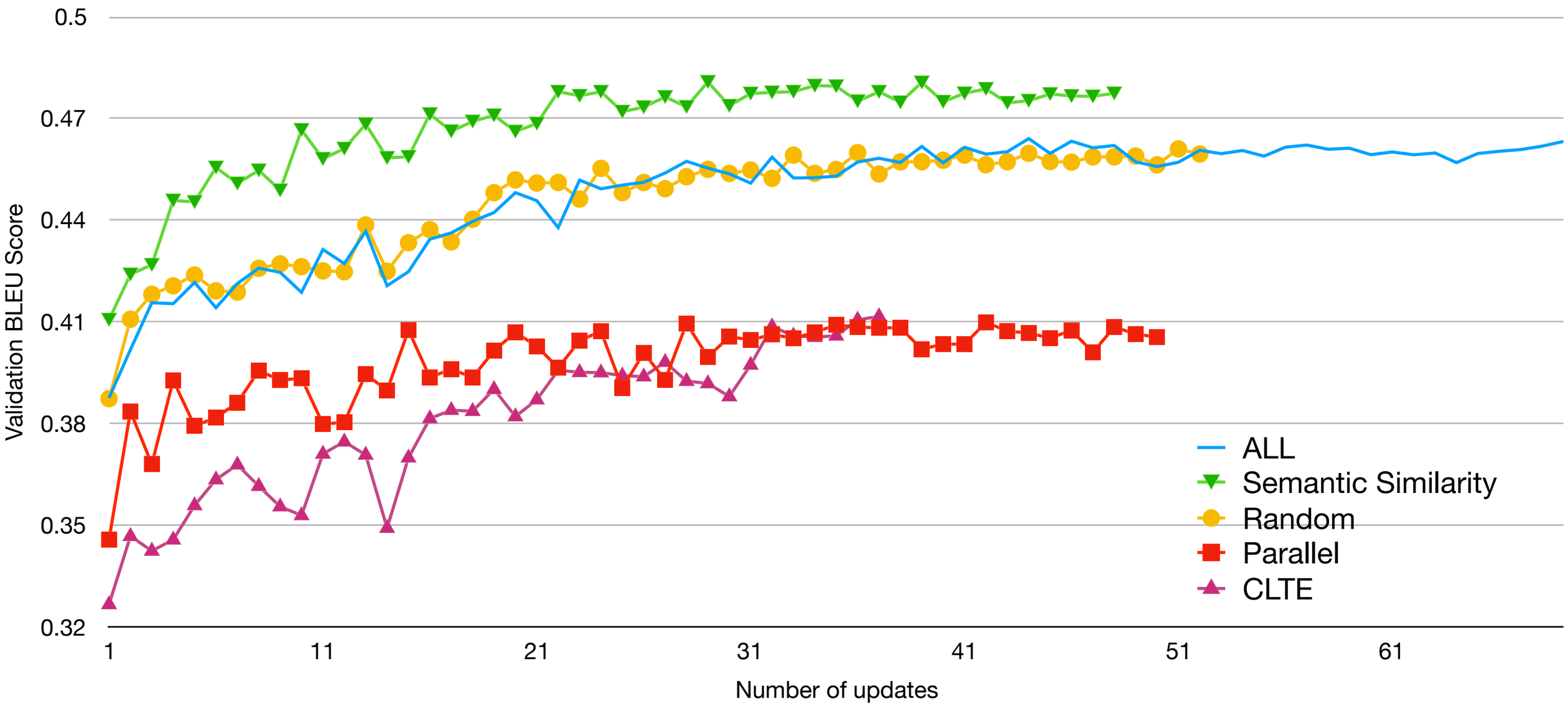
Faster and Better NMT Training



Faster and Better NMT Training

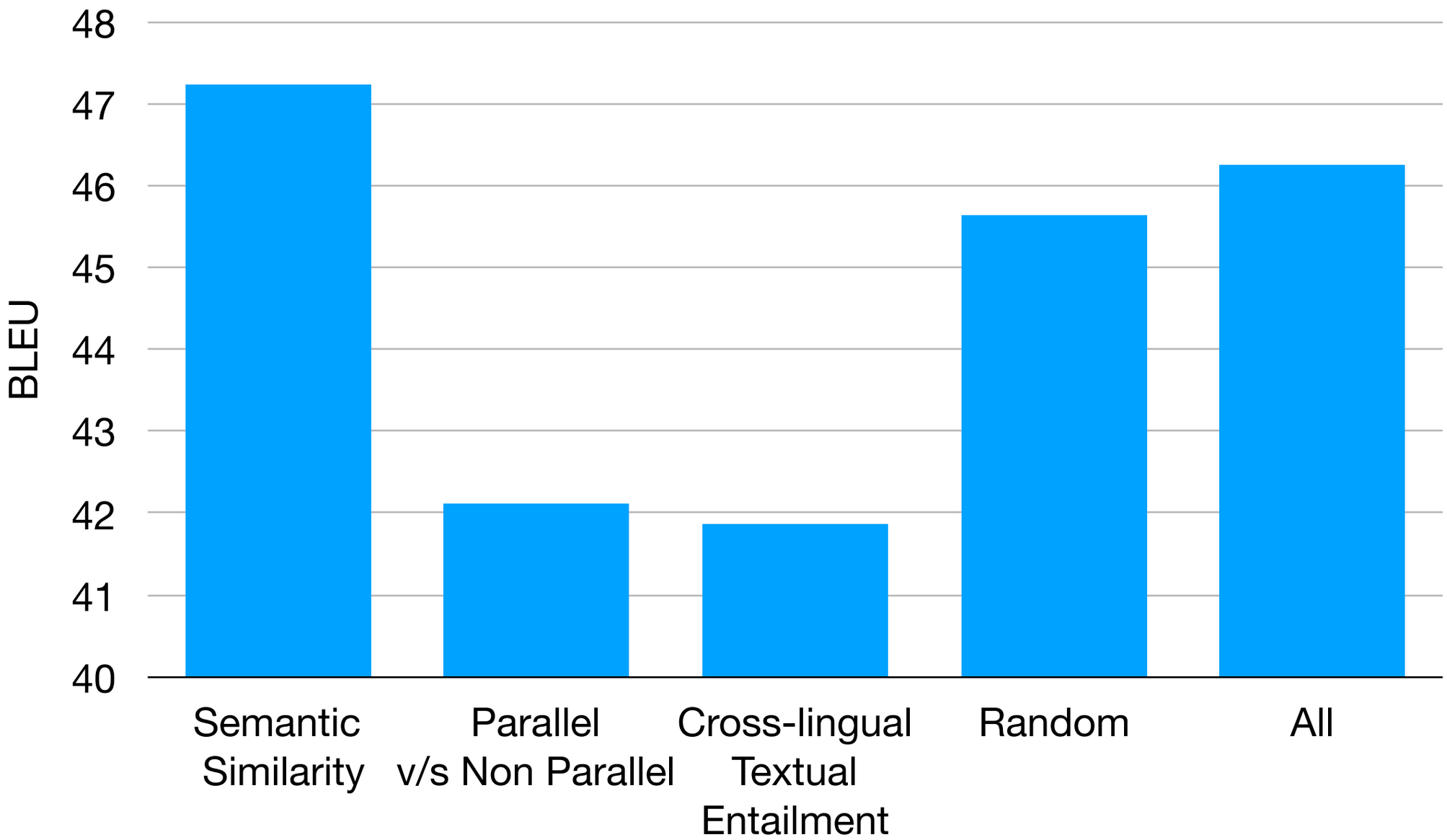


Faster and Better NMT Training

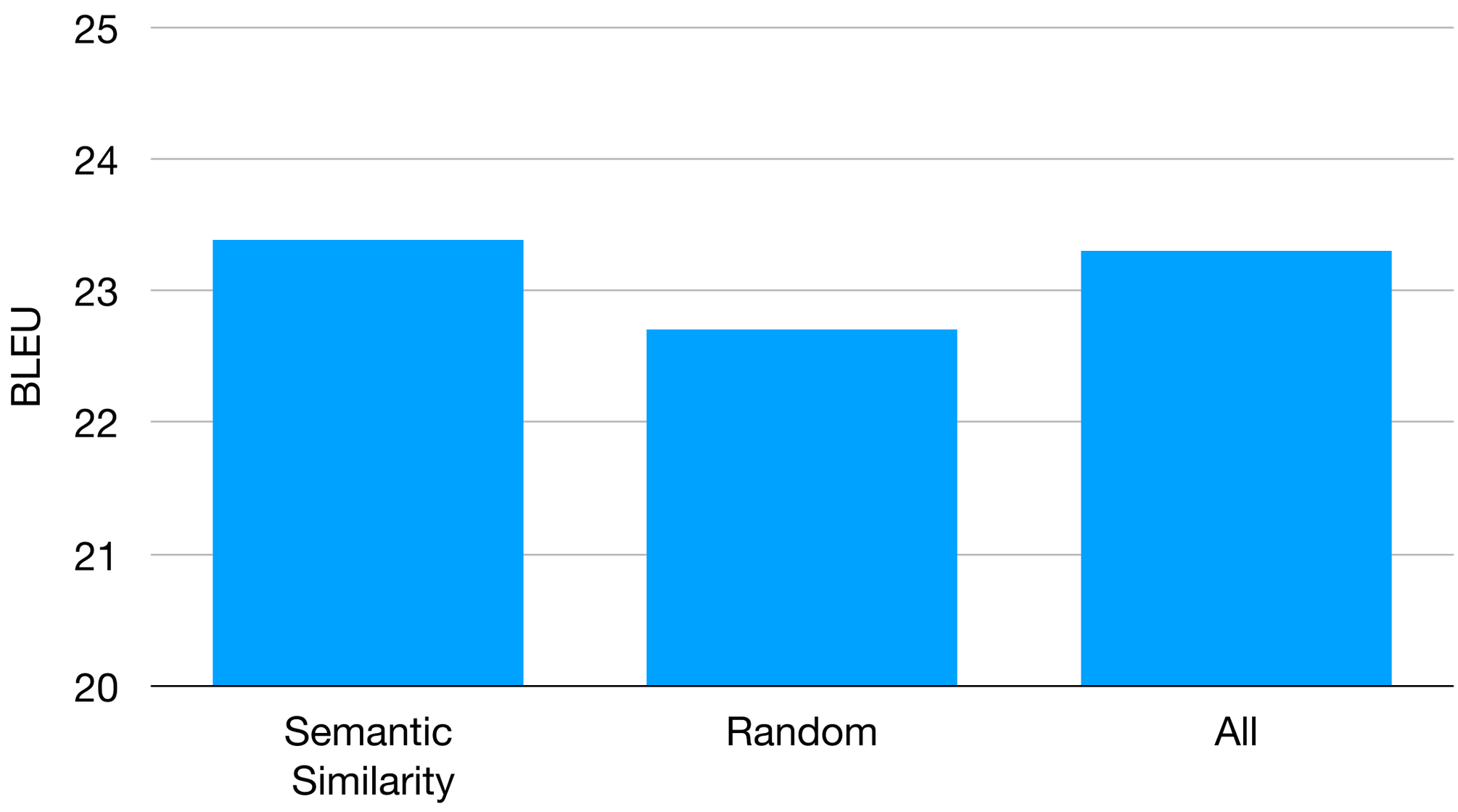


Better translation quality, with only 14M sentence pairs (instead of 28M)

Decoding Results - Fr-En



Divergences also impact low-resource NMT



Analysis

- **Intrinsic experiments**

  ➡ Semantic Similarity model has **low false positives** on non-divergent class

Someone wanted to cook bratwurst	Vous vouliez des saucisses grillées `you wanted some grilled sausages`

Semantic Similarity prediction	**Divergent**	
Parallel v/s non parallel prediction	**Non - divergent**	

Analysis

- **Intrinsic experiments**

  ➡ Semantic Similarity model has **low false positives** on non-divergent class

- **Neural MT experiments**

  ➡ Other models fail due to **lack of adequacy**

Analysis

- **Intrinsic experiments**

  ➡ Semantic Similarity model has **low false positives** on non-divergent class

- **Neural MT experiments**

  ➡ Other models fail due to **lack of adequacy**

  - Model trained via CLTE under-translates by dropping segments

Analysis

- **Intrinsic experiments**

  ➡ Semantic Similarity model has **low false positives** on non-divergent class

- **Neural MT experiments**

  ➡ Other models fail due to **lack of adequacy**

  - Model trained via CLTE under-translates by dropping segments

Source	Reference	Translation
he's a very impressive man and still goes out to do digs	c'est un homme très impressionnant **et il fait encore des fouilles.**	c'est un homme très impressionnant.
when the Heat first won.	lorsque les Heat ont gagné **pour la première fois.**	quand le Heat a gagné.

Analysis

- **Intrinsic experiments**

  ➡ Semantic Similarity model has **low false positives** on non-divergent class

- **Neural MT experiments**

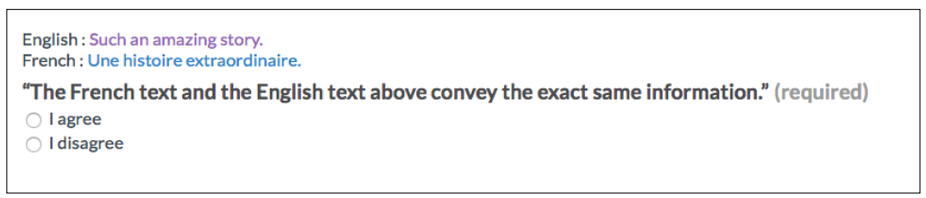
  ➡ Other models fail due to **lack of adequacy**

  - Model trained via CLTE under-translates by dropping segments
  - Parallel v/s non-parallel model generates garbage

Key Findings : Semantic Divergences ..

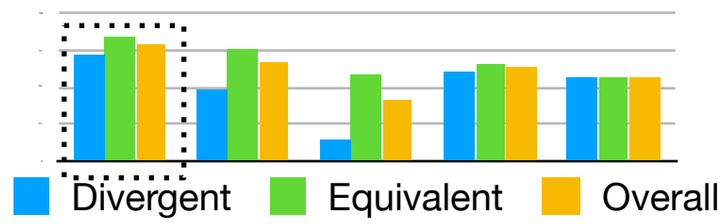
- Are common in parallel data

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)



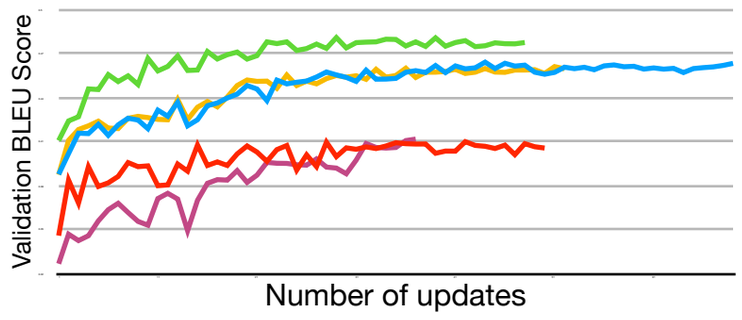
- Can be detected without manual annotations using a deep model of bilingual similarity

  ➡ **80 F1** on a crowdsourced sample



- **Have a measurable impact on NMT training**

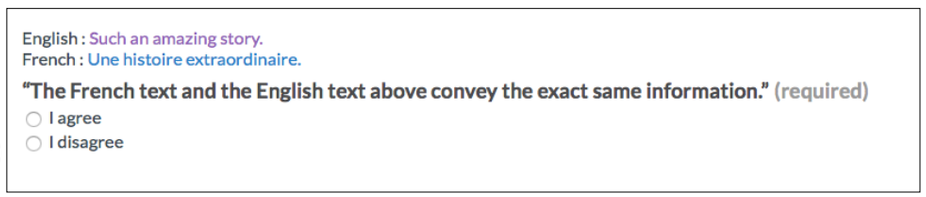
  ➡ Discard most divergent examples yields better BLEU, in less training time



Key Findings : Semantic Divergences ..

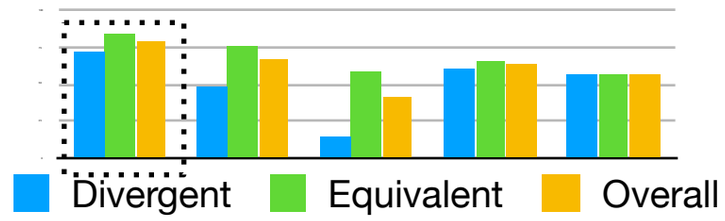
- Are common in parallel data

  ➡ **~40%** in En-Fr (OpenSubtitles and CommonCrawl)



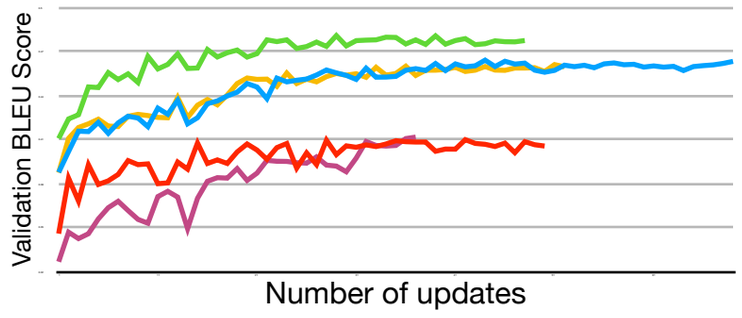
- Can be detected without manual annotations using a deep model of bilingual similarity

  ➡ **80 F1** on a crowdsourced sample



- Have a measurable impact on NMT training

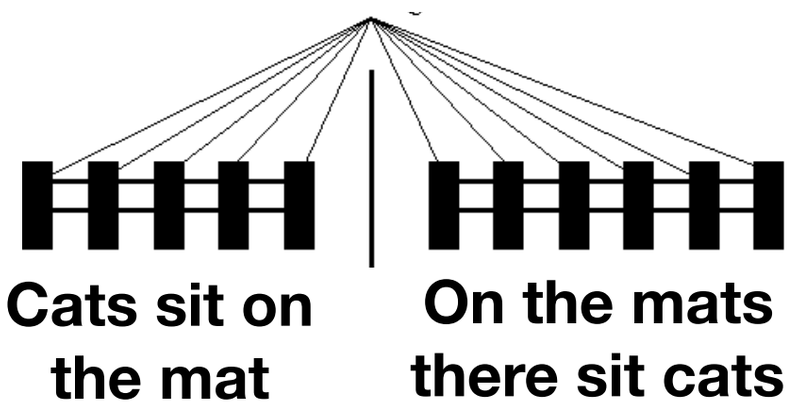
  ➡ Discard most divergent examples yields better BLEU, in less training time



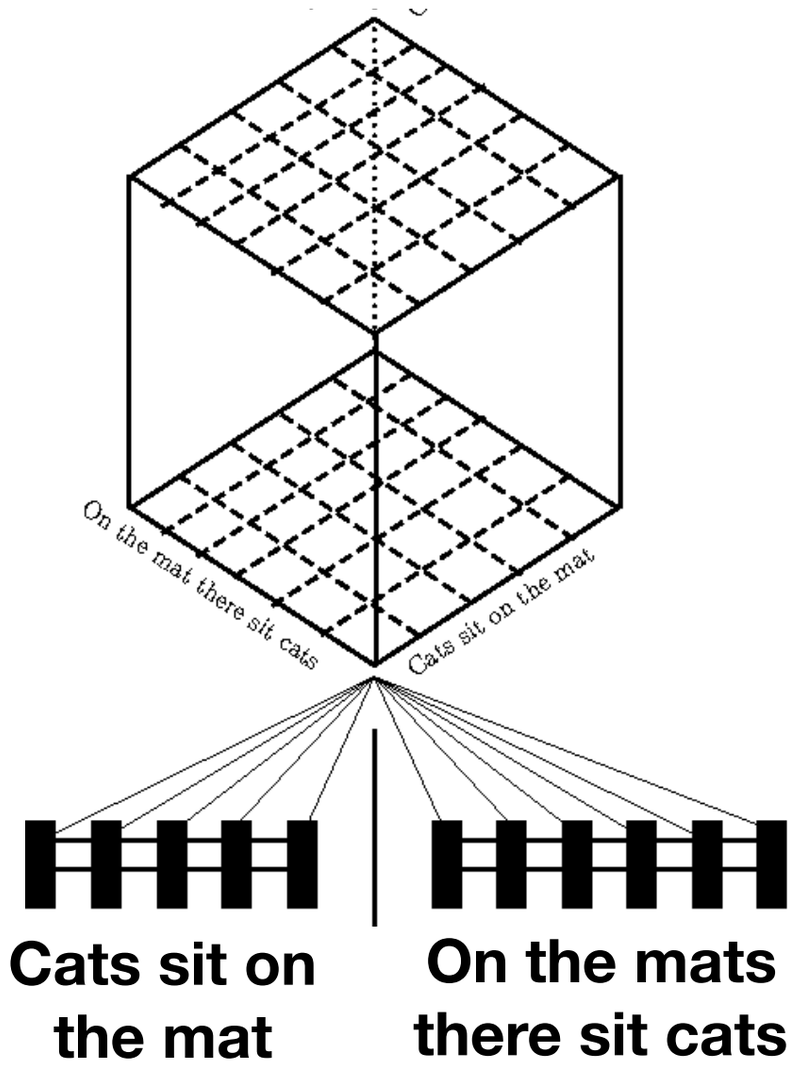
Backup

Cats sit on the mat

On the mats there sit cats

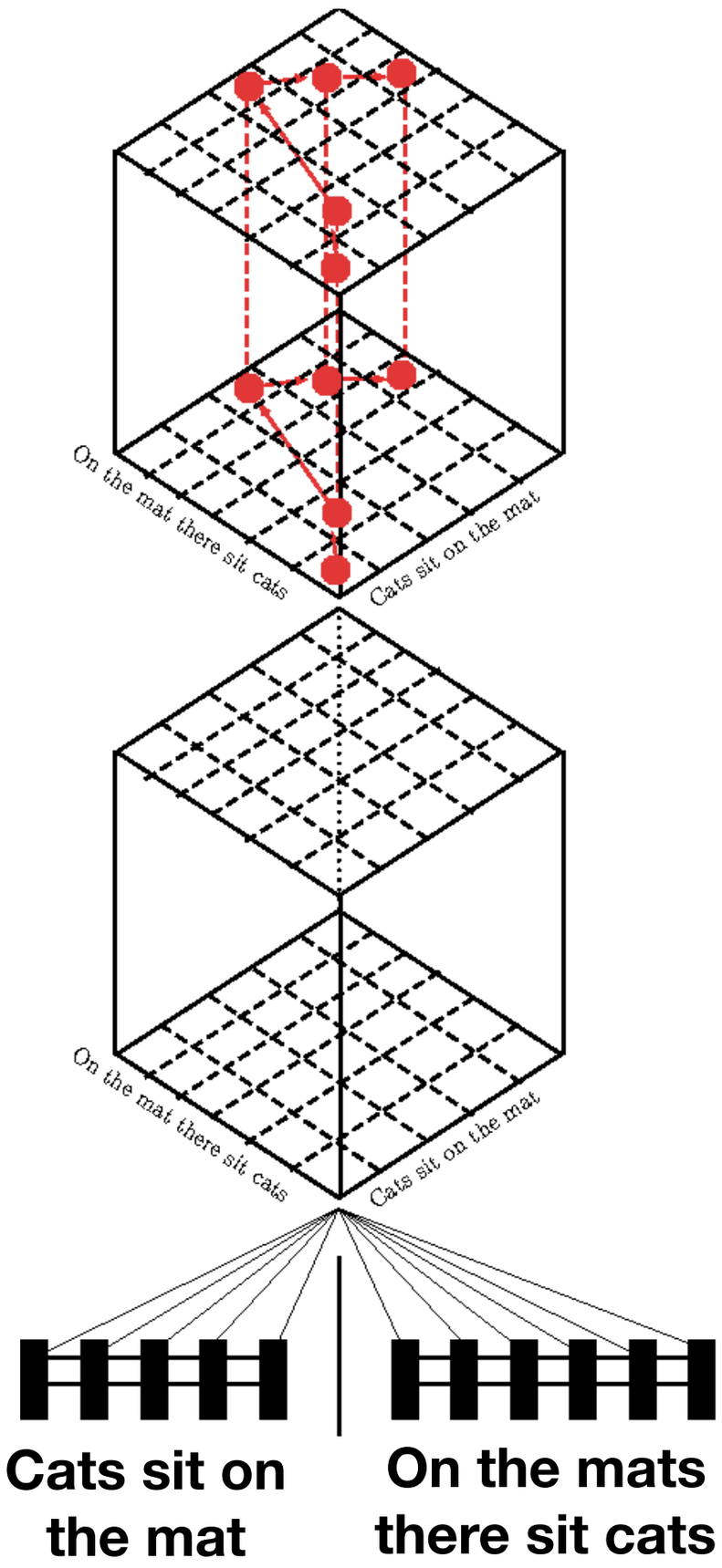


1. BiLSTM to represent words in context



2. Tensor of similarities (similarity cube)

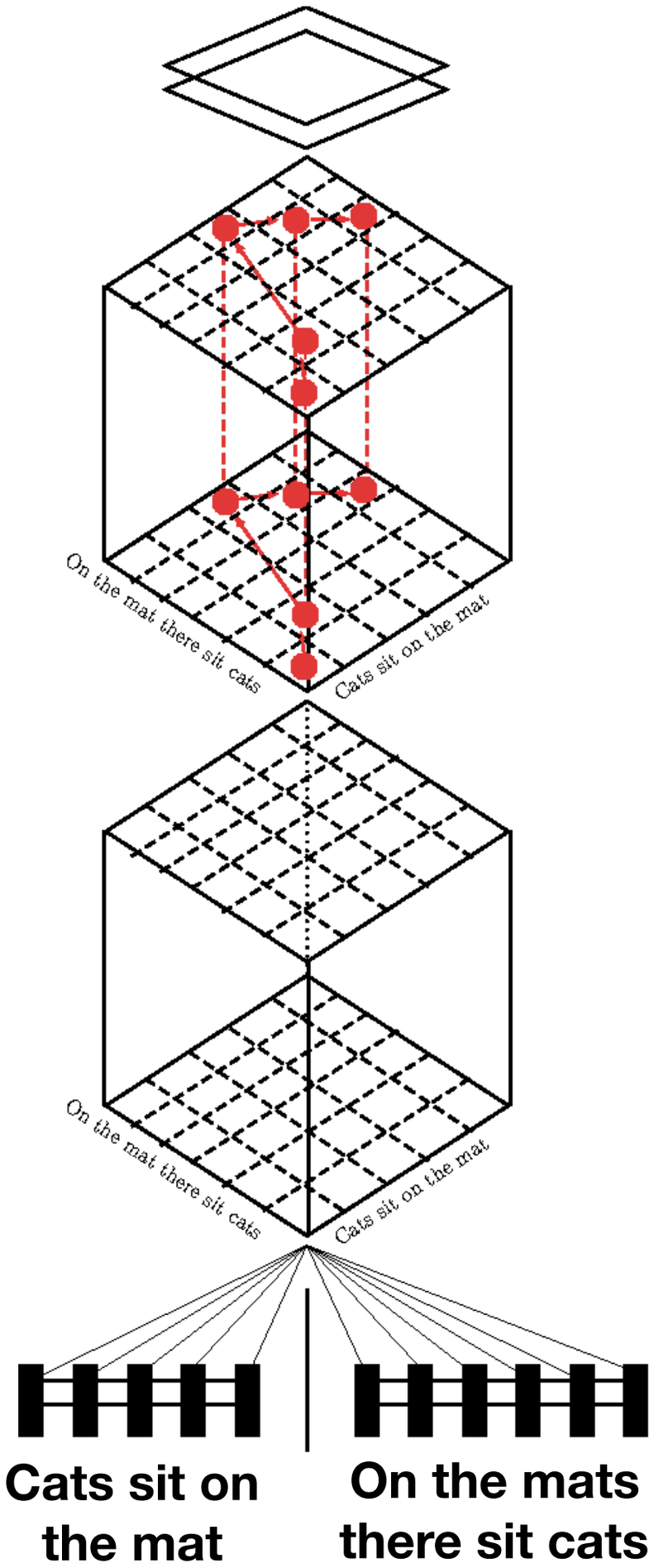
1. BiLSTM to represent words in context



3. Identifying most salient word pairs (Focus cube)

2. Tensor of similarities (similarity cube)

1. BiLSTM to represent words in context



4. Deep CNN to extract features and classify

3. Identifying most salient word pairs (Focus cube)

2. Tensor of similarities (similarity cube)

1. BiLSTM to represent words in context

Semantic Divergences in Parallel Data

- Parallel sentences == sentences that are translations of each other
  - ➡ Do they always convey the same meaning?

I don’t know what I’m going to do.	J’en sais rien *(I do not know.)*
Someone wanted to cook bratwurst	Vous vouliez des saucisses grillées *(you wanted some grilled sausages)*

- Semantic divergences can arise due to :
  - ➡ Bad translations
  - ➡ Alignment Noise (Munteanu and Marcu, 2004, 2005)
  - ➡ Subtle language-specific nuances (lexical choice, discourse effects, etc.)

Semantic Divergences in Parallel Data

- Parallel sentences == sentences that are translations of each other
  - ➡ Do they always convey the same meaning?

I don’t know what I’m going to do.	J’en sais rien I do not know.
Someone wanted to cook bratwurst	Vous vouliez des saucisses grillées you wanted some grilled

- Semantic divergences can arise due to :
  - ➡ Bad translations
  - ➡ Alignment Noise (Munteanu and Marcu, 2004, 2005)
  - ➡ Subtle language-specific nuances (lexical choice, discourse effects, etc.)

- Also differ from typological divergences (Dorr, 1994)

Detecting Divergences can help MT

- Statistical MT relatively robust to noise, e.g. sentence mis-alignments (Goutte et al., 2012)

Detecting Divergences can help MT

- Statistical MT relatively robust to noise, e.g. sentence mis-alignments (Goutte et al., 2012)

- Neural MT more sensitive to the nature of examples (Chen et al., 2016, Belinkov and Bisk, 2018)

  ➡ Filtering out divergent pairs can assist NMT training

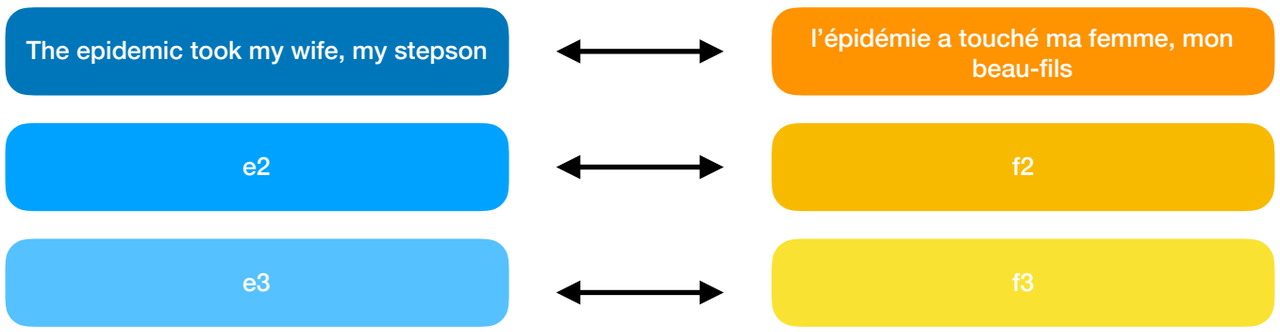
Detecting Divergences can help MT

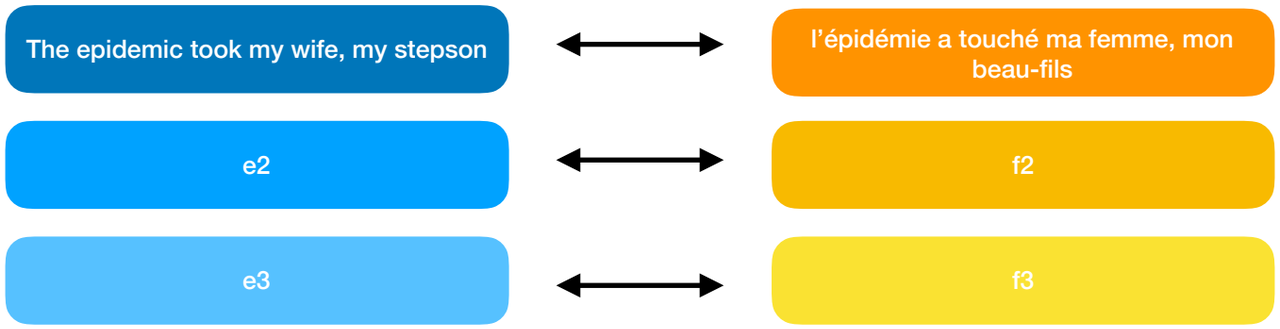
- Statistical MT relatively robust to noise, e.g. sentence mis-alignments (Goutte et al., 2012)

- Neural MT more sensitive to the nature of examples (Chen et al., 2016, Belinkov and Bisk, 2018)
  - ➡ Filtering out divergent pairs can assist NMT training

- Other NLP tasks can also benefit
  - ➡ Bilingual embeddings, cross-lingual projections

Noisy Synthetic Supervision





(Munteanu and Marcu 2005)