Cross-Lingual Text Classification With Minimal Resources By Transferring a Sparse Teacher

Giannis Karamanolakis, Daniel Hsu, Luis Gravano
Department of Computer Science, Columbia University
{gkaraman, djhsu, gravano}@cs.columbia.edu

COLUMBIA UNIVERSITY

Document Classification Beyond English: A Labeled Data Bottleneck

- Most NLP techniques/datasets are developed in English
- 7,000 living languages (~4,000 written)
- Our focus: multilingual document classification (e.g., emergency detection in Uyghur)

source: http://endangeredlanguages.com/

Document Classification Beyond English: A Labeled Data Bottleneck

- Most NLP techniques/datasets are developed in English
- 7,000 living languages (~4,000 written)
- Our focus: multilingual document classification (e.g., emergency detection in Uyghur)
    - **Issue:** expensive to obtain labeled documents
    - **Cross-lingual classification:** use labeled documents from a source language

source: http://endangeredlanguages.com/

Cross-Lingual Text Classification: Approaches & Resources

- Challenge: how to bridge the source and target languages?

“We needed medical supplies”

Source Language

??

“بىزگە داۋالاش بۇيۇملىرى لازىم”

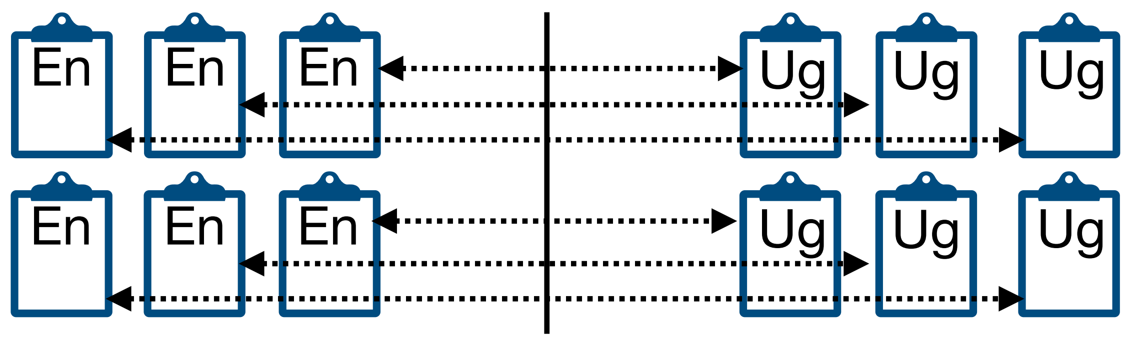
Target Language

Cross-lingual resources required!

Cross-Lingual Text Classification: Approaches & Resources

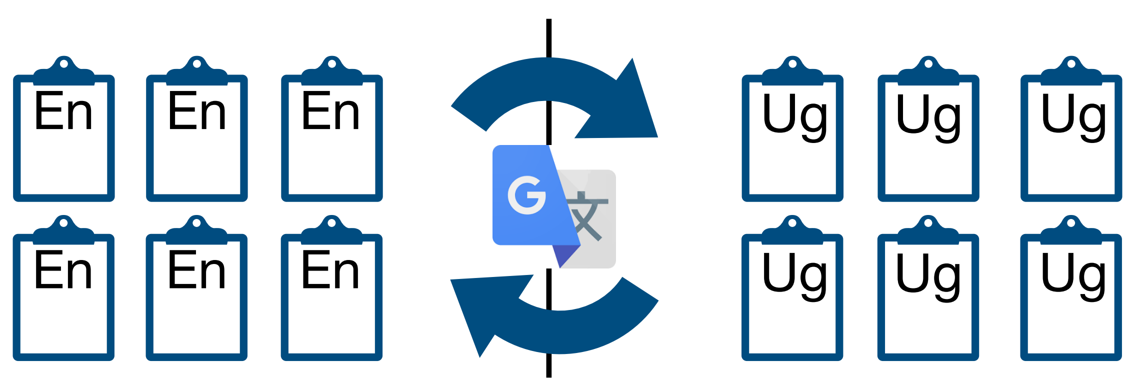
- Challenge: how to bridge the source and target languages?
    - Approach 1: Transfer supervision across languages

Parallel Corpora



or

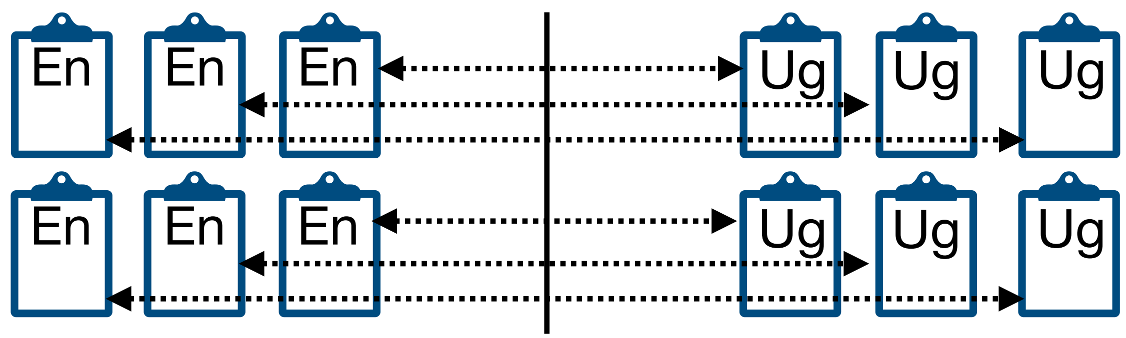
Machine Translation



Cross-Lingual Text Classification: Approaches & Resources

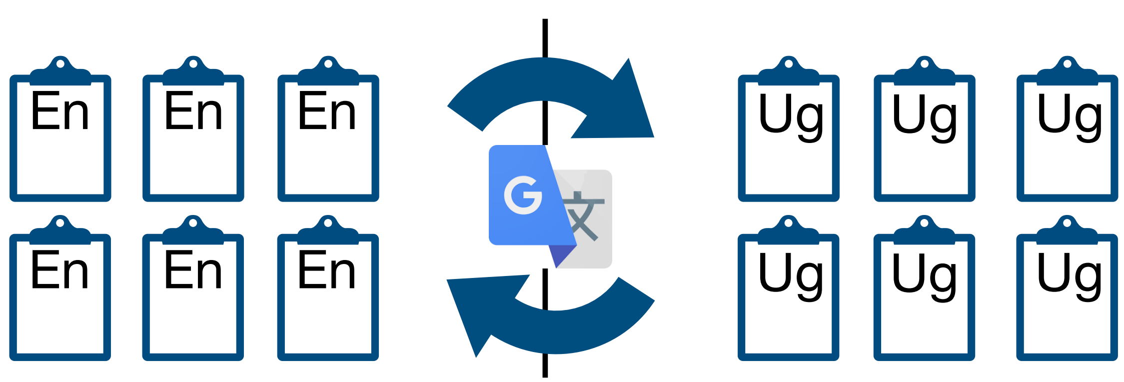
- Challenge: how to bridge the source and target languages?
  - Approach 1: Transfer supervision across languages **(-) expensive**

Parallel Corpora



or

Machine Translation



Google Translate is available for 103/4,000 languages

Cross-Lingual Text Classification: Approaches & Resources

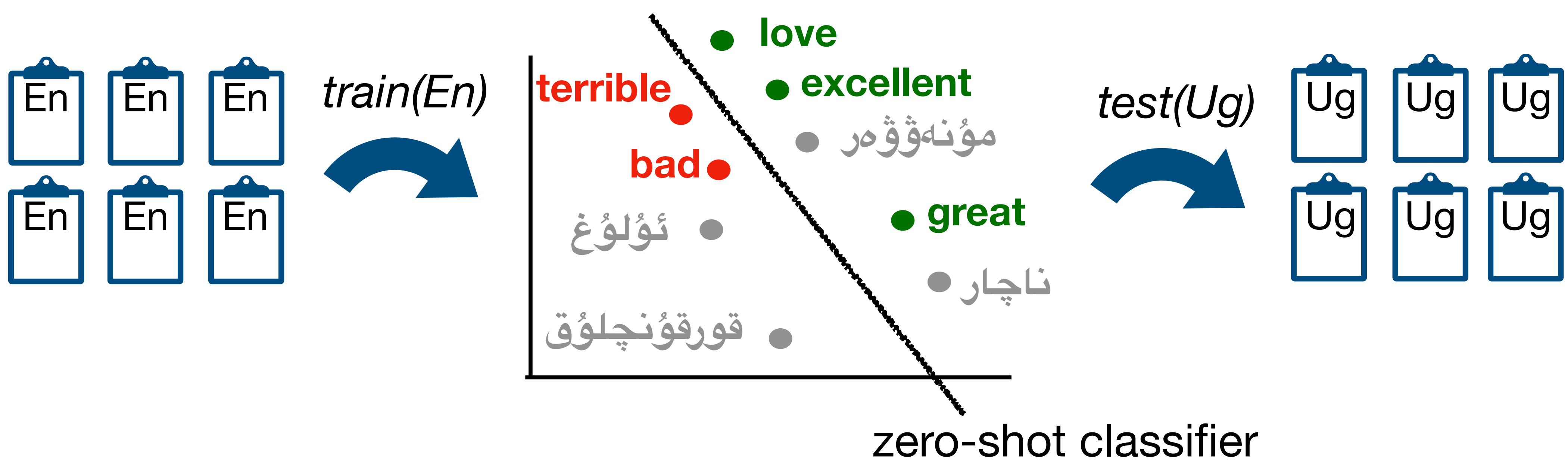
- Challenge: how to bridge the source and target languages?
  - Approach 1: Transfer supervision across languages **(-) expensive**
  - Approach 2: Train zero-shot classifiers

Pre-trained Cross-lingual Embeddings / Multilingual Language Models

Cross-Lingual Text Classification: Approaches & Resources

- Challenge: how to bridge the source and target languages?
  - Approach 1: Transfer supervision across languages **(-) expensive**
  - Approach 2: Train zero-shot classifiers

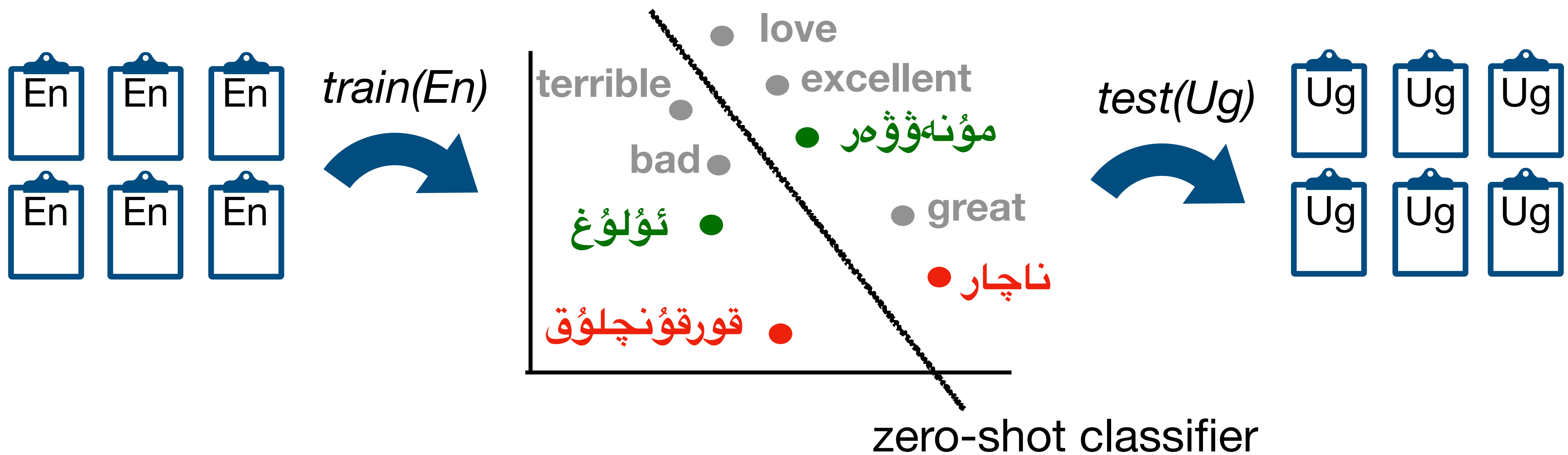
Pre-trained Cross-lingual Embeddings / Multilingual Language Models



Cross-Lingual Text Classification: Approaches & Resources

- Challenge: how to bridge the source and target languages?
  - Approach 1: Transfer supervision across languages **(-) expensive**
  - Approach 2: Train zero-shot classifiers

Pre-trained Cross-lingual Embeddings / Multilingual Language Models



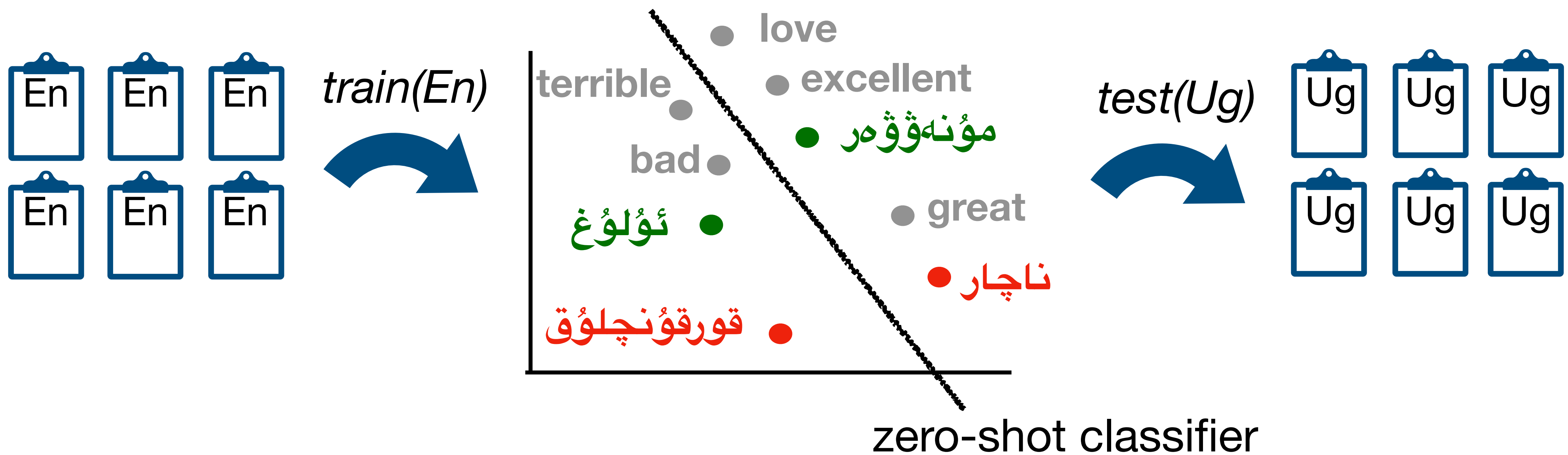
(-) Target-language documents are not considered during training

- May not capture patterns specific to the target language or task

Cross-Lingual Text Classification: Approaches & Resources

- Challenge: how to bridge the source and target languages?
  - Approach 1: Transfer supervision across languages **(-) expensive**
  - Approach 2: Train zero-shot classifiers

Pre-trained Cross-lingual Embeddings / Multilingual Language Models



(-) Target-language documents are not considered during training
- May not capture patterns specific to the target language or task

(-) High-quality representations are not always available
- Multilingual BERT available for only 104 out of 4,000 languages
- High-coverage bilingual dictionaries not available for all languages

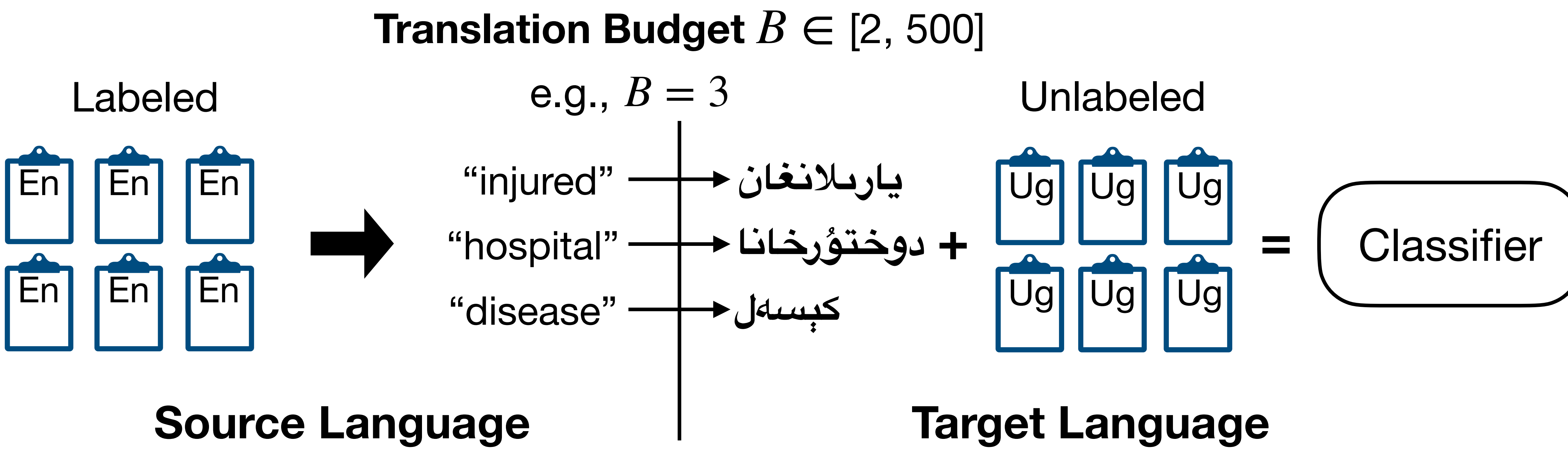
Cross-Lingual Text Classification: Approaches & Resources

- Challenge: how to bridge the source and target languages?
  - Approach 1: Transfer supervision across languages **(-) expensive**
  - Approach 2: Train zero-shot classifiers **(-) not effective / not available**
  - Our approach: Transfer **weak** supervision using **minimal** resources

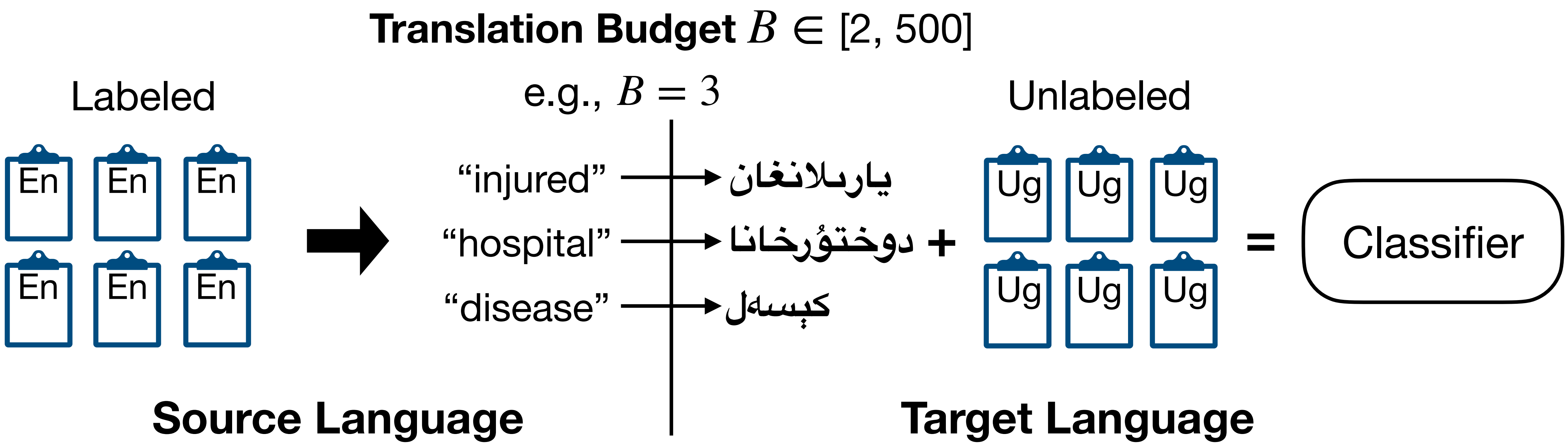
(+) Does **not** require parallel corpora / machine translation / multilingual representations

(+) Has **robust** performance across **18 diverse languages** and 4 tasks

We Transfer Weak Supervision Using Minimal Resources



We Transfer Weak Supervision Using Minimal Resources



- **Our contributions:**

1. Present a method for cross-lingual transfer under a **limited translation budget**
2. Show how to train **any** target classifier **without labeled target documents**
3. Show the benefits of generating weak supervision in **18 diverse languages**

Outline

1. Intro: Cross-Lingual Text Classification

2. Our Approach: Cross-Lingual Teacher-Student (CLTS)

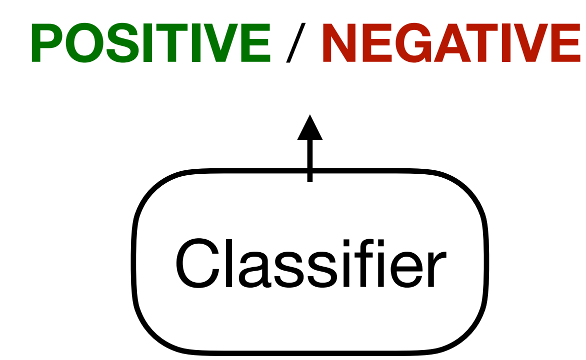
3. Experiments in 18 Languages

4. Conclusions

Cross-Lingual Transfer Under Limited Translation Budget

- Goal: train a target classifier given
  - **Labeled** documents in the **source** language
  - **Unlabeled** documents in the **target** language
  - **Budget** for up to B word translations

$B = 2$



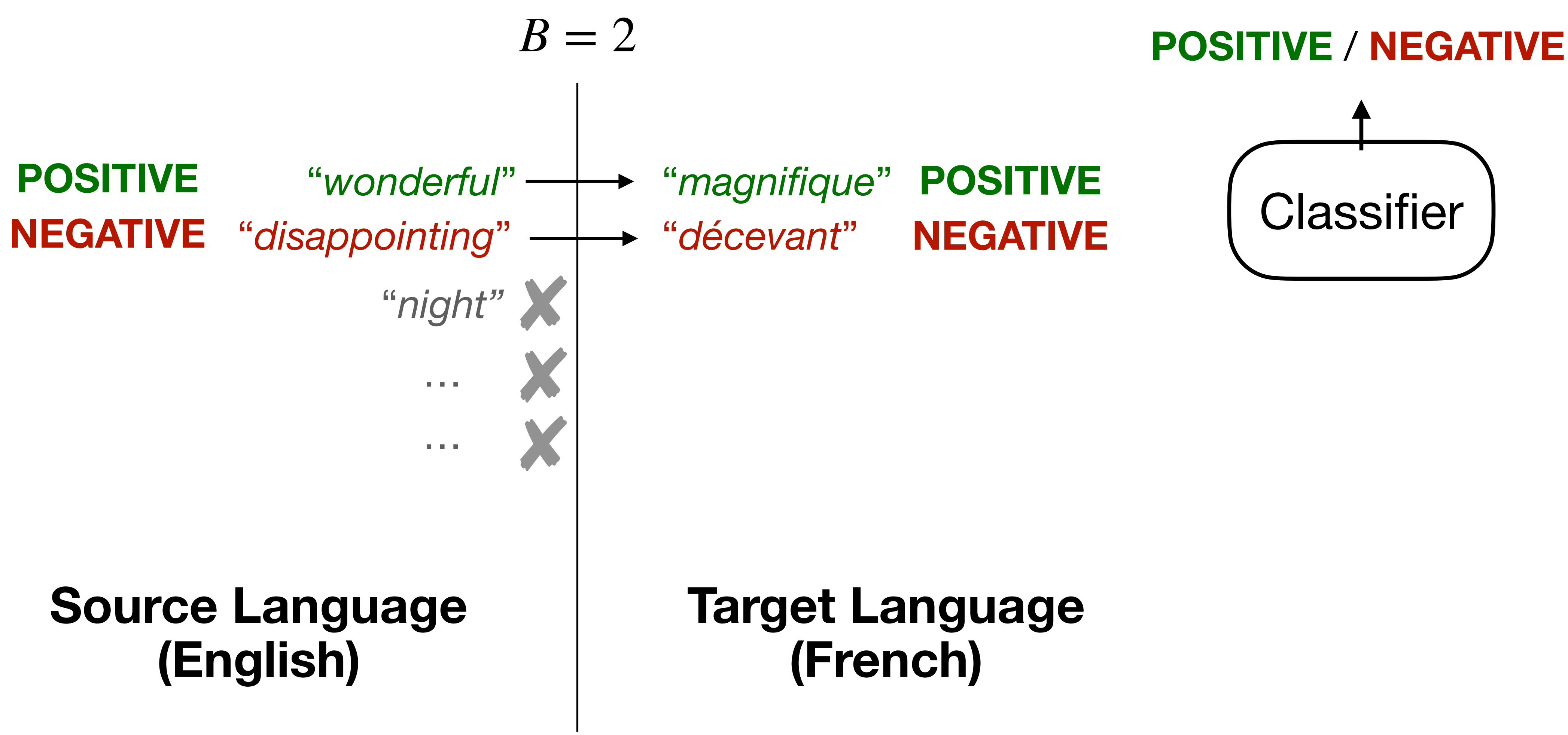
Source Language (English)

Target Language (French)

Cross-Lingual Transfer Under Limited Translation Budget

- Goal: train a target classifier given
  - **Labeled** documents in the **source** language
  - **Unlabeled** documents in the **target** language
  - **Budget** for up to B word translations

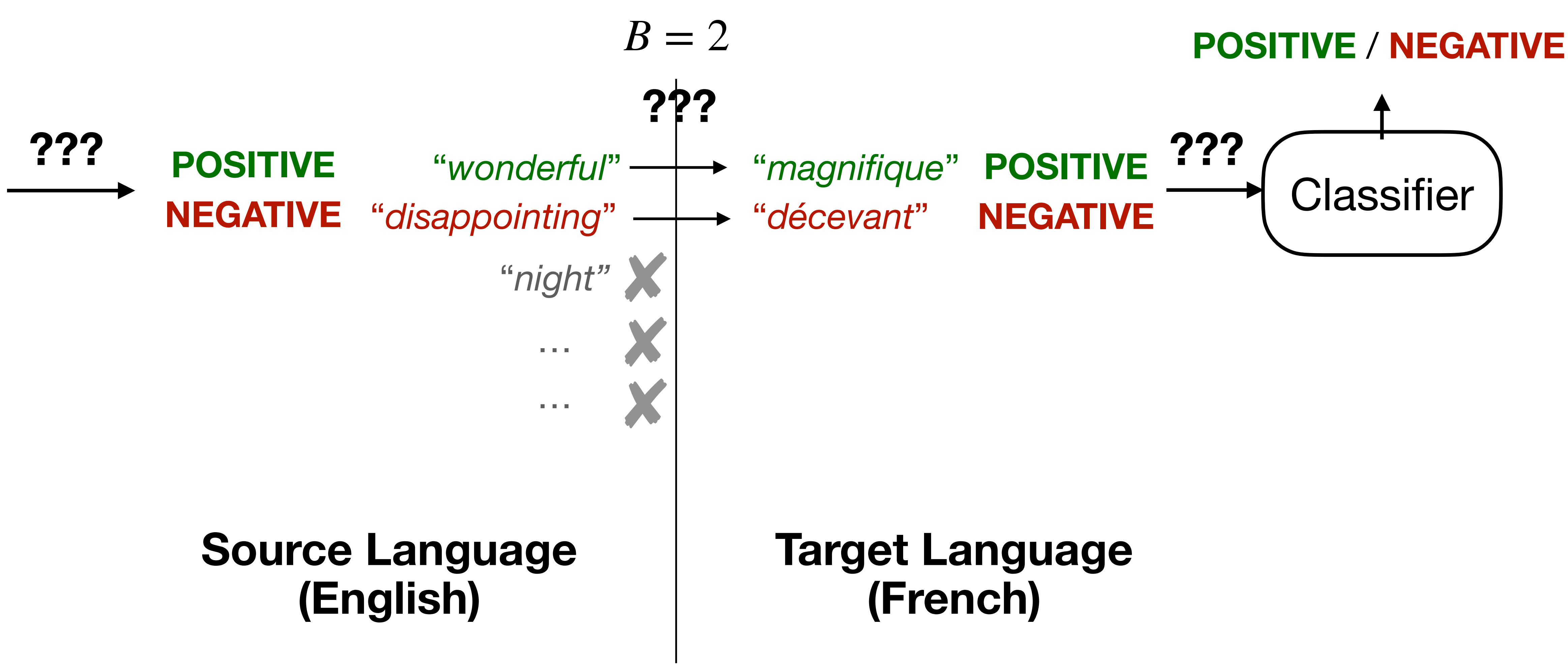
- Our idea: transfer only the most **indicative keywords** (seed words)



Cross-Lingual Transfer Under Limited Translation Budget

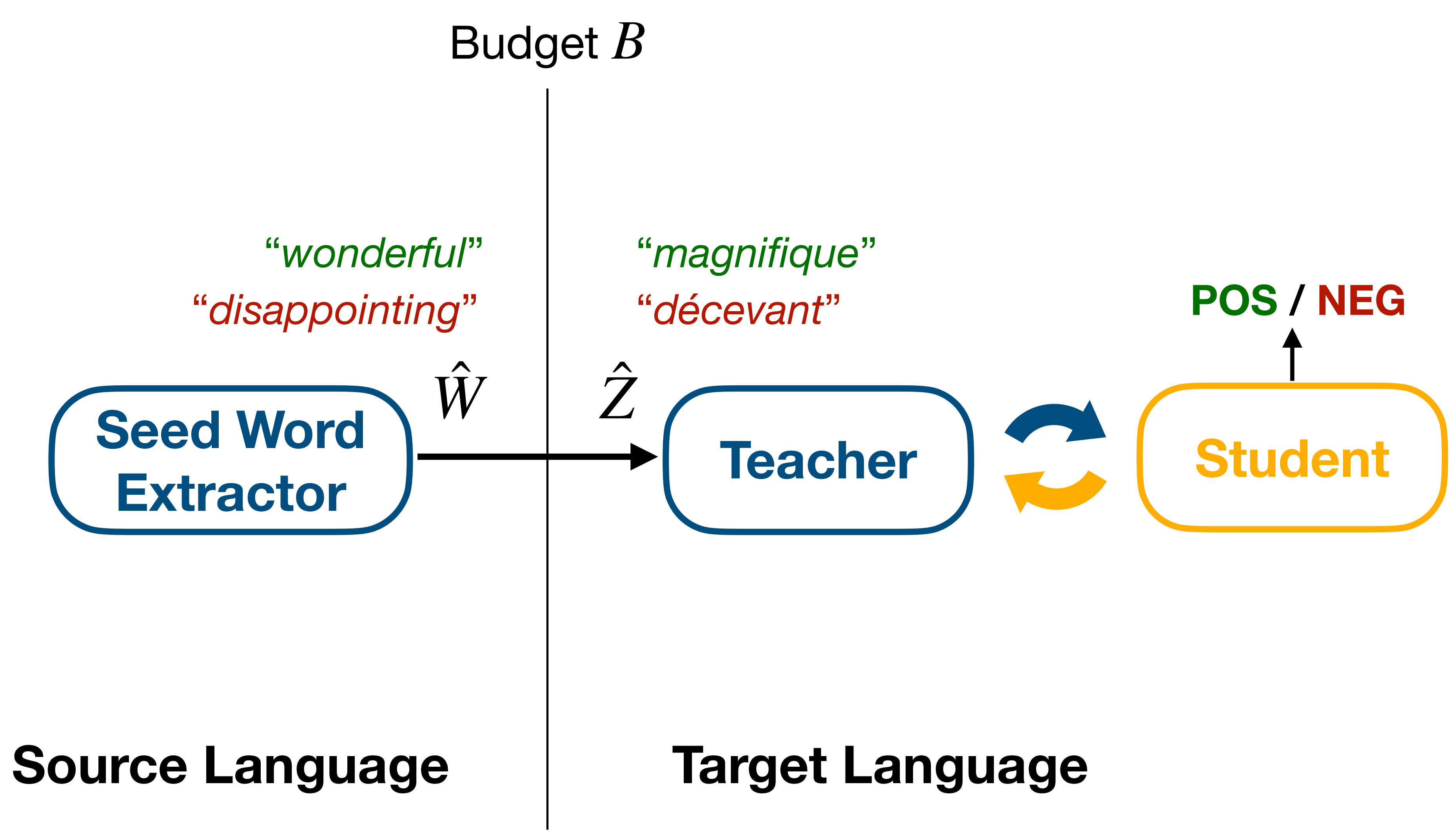
- Goal: train a target classifier given
  - **Labeled** documents in the **source** language
  - **Unlabeled** documents in the **target** language
  - **Budget** for up to B word translations

- Our idea: transfer only the most **indicative keywords** (seed words)



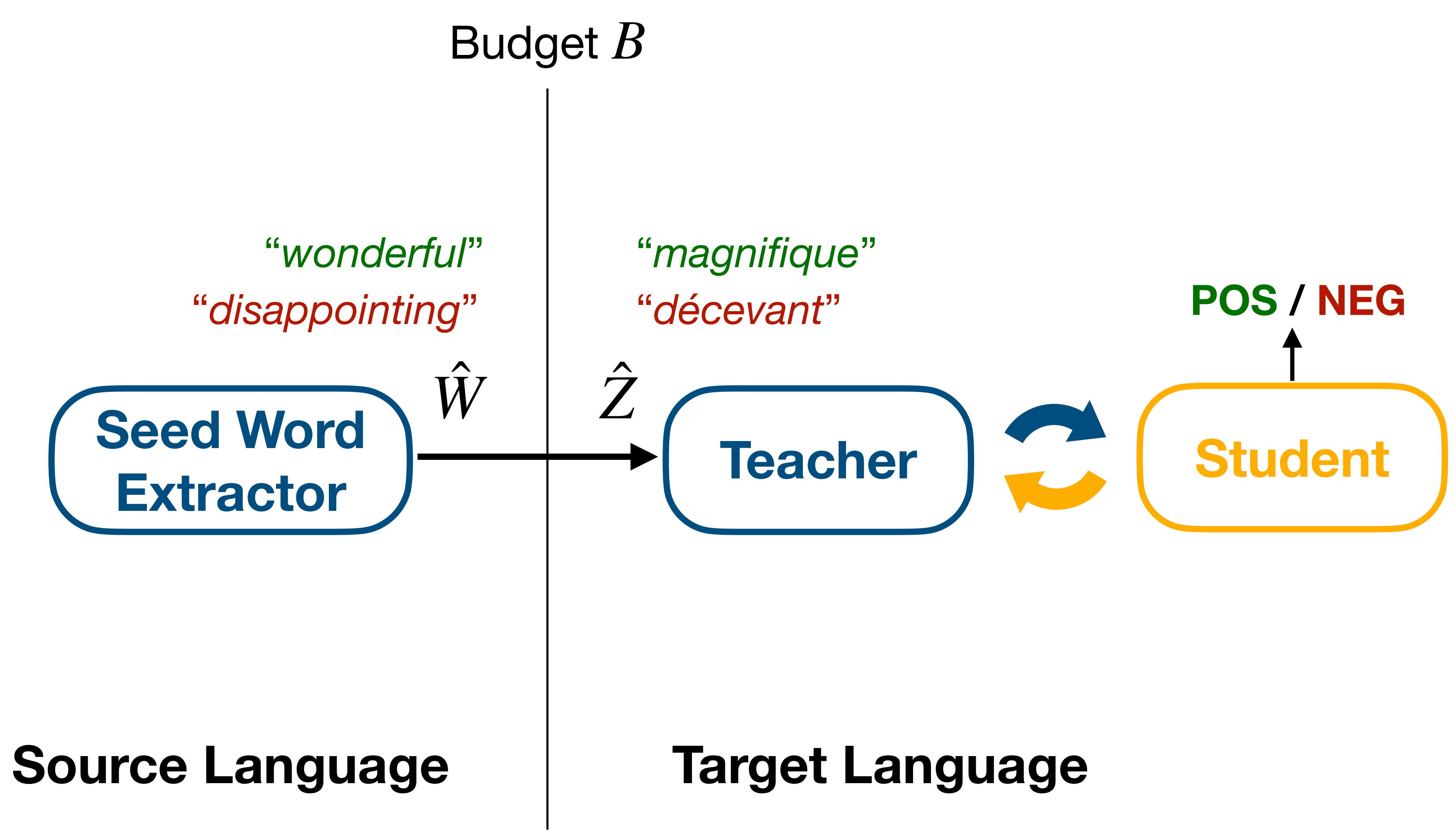
Cross-Lingual Teacher-Student (CLTS)

1. Seed-word extraction in the **source** language
2. Cross-lingual seed **weight transfer**
3. Teacher-Student co-training in the **target** language



Cross-Lingual Teacher-Student (CLTS)

1. **Seed-word extraction in the source language**
2. Cross-lingual seed weight transfer
3. Teacher-Student co-training in the **target** language

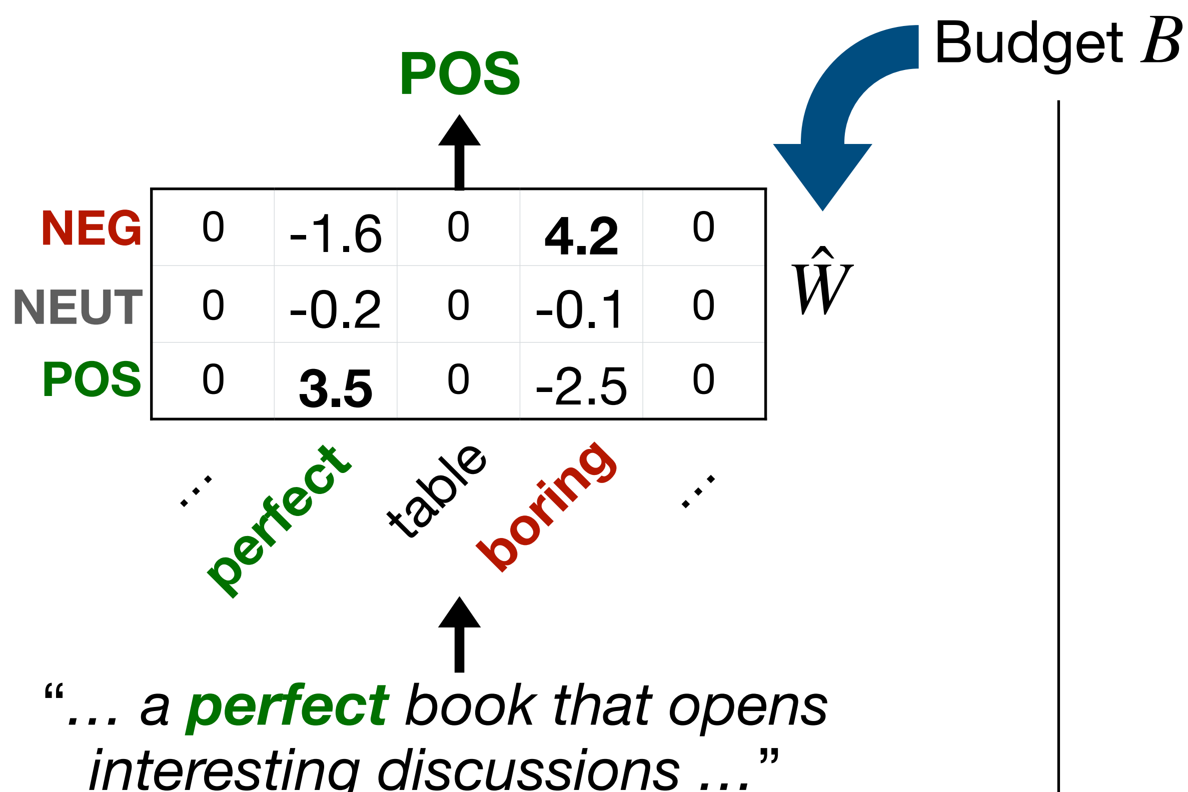


Training a Sparse Classifier in the Source Language

1. Seed-word extraction in the source language

- Extract B seed words from the weight matrix $\hat{W}$ of a classifier
- Use B as sparsity regularizer **during** training

Sparse Classifier



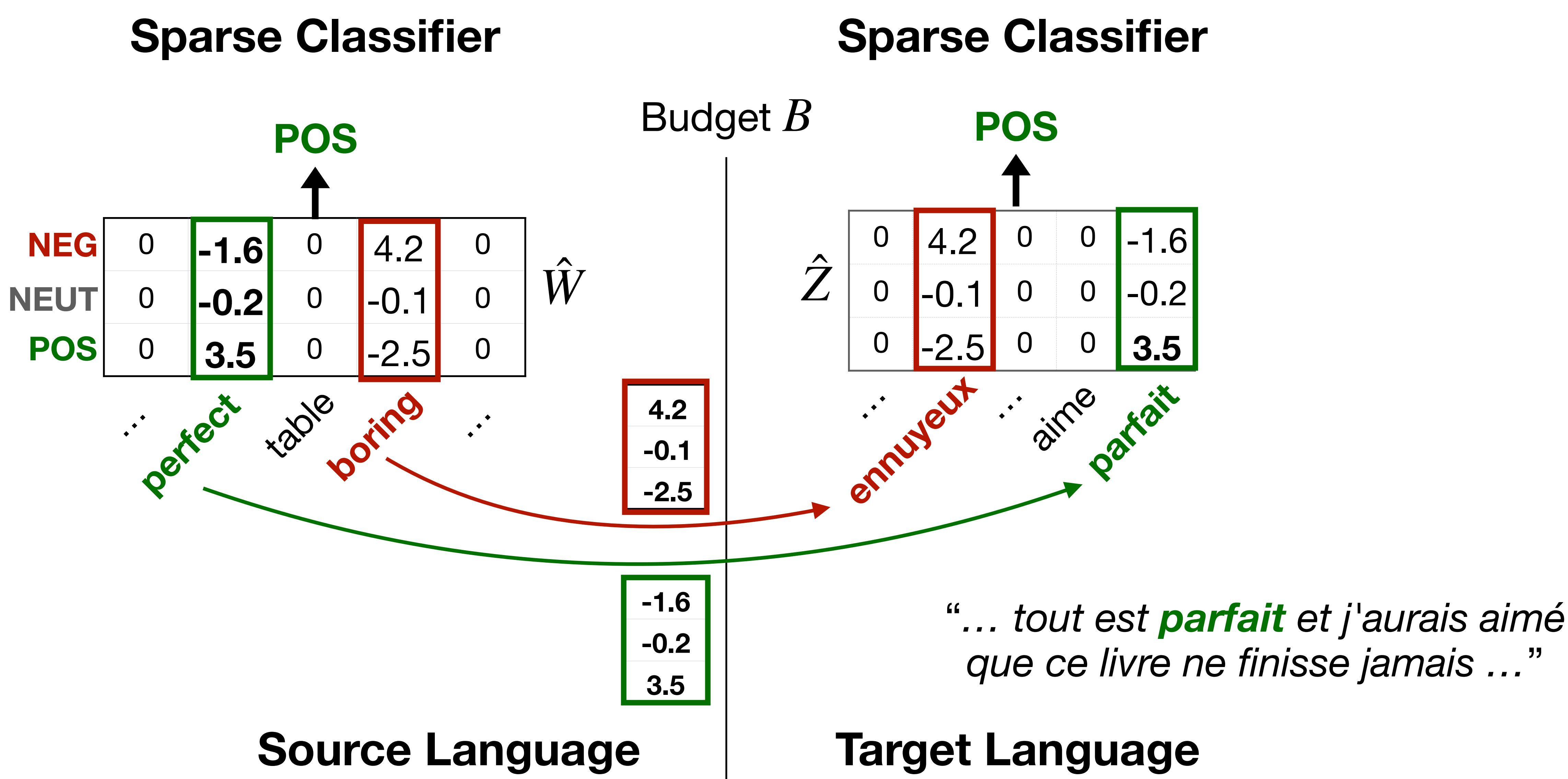
Source Language | **Target Language**

Transferring the Sparse Classifier Across Languages

1. Seed-word extraction in the source language

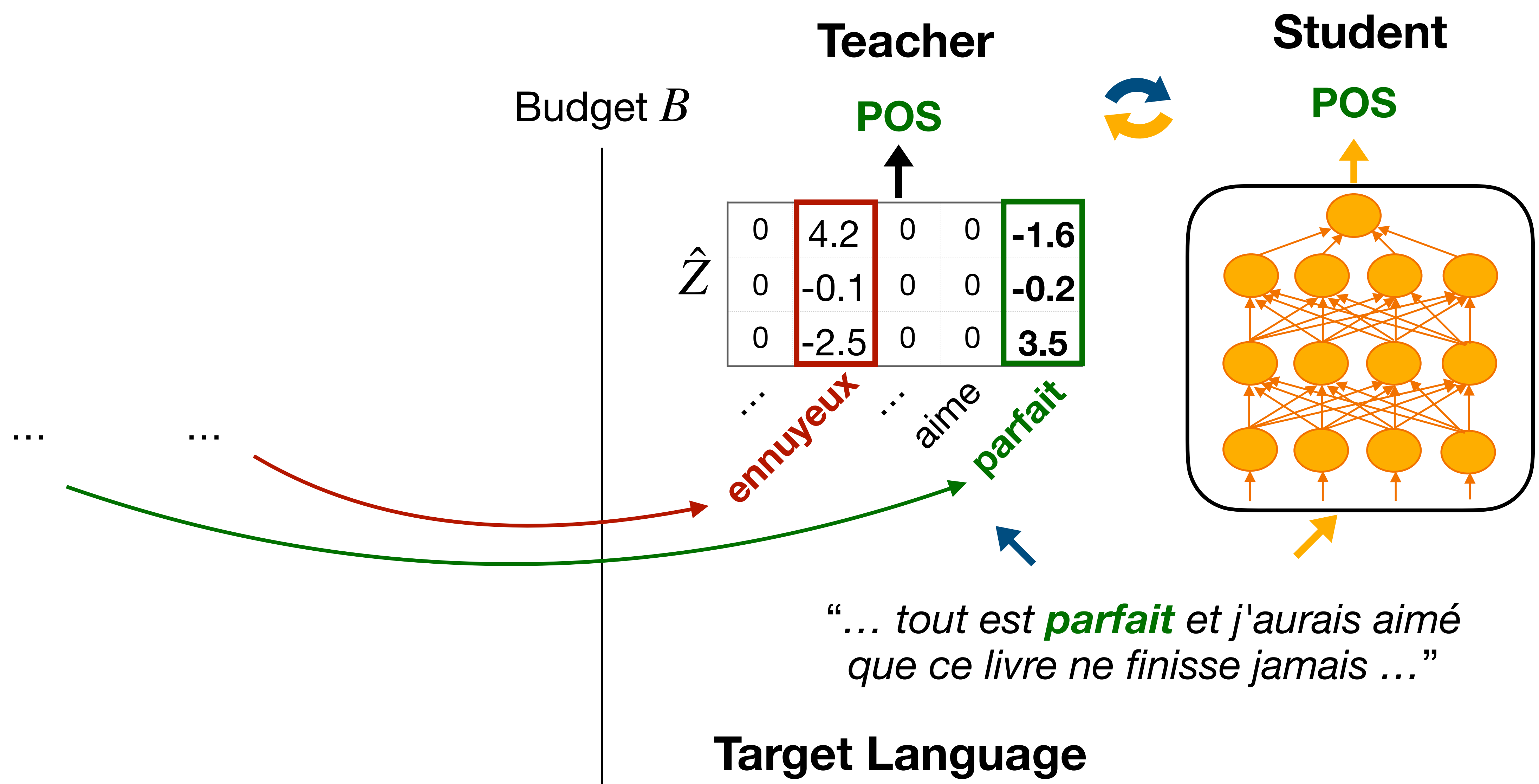
2. Cross-lingual seed weight transfer

- Obtain translations for the B seed words and transfer their weights
- Initialize target classifier based on the translated seed words



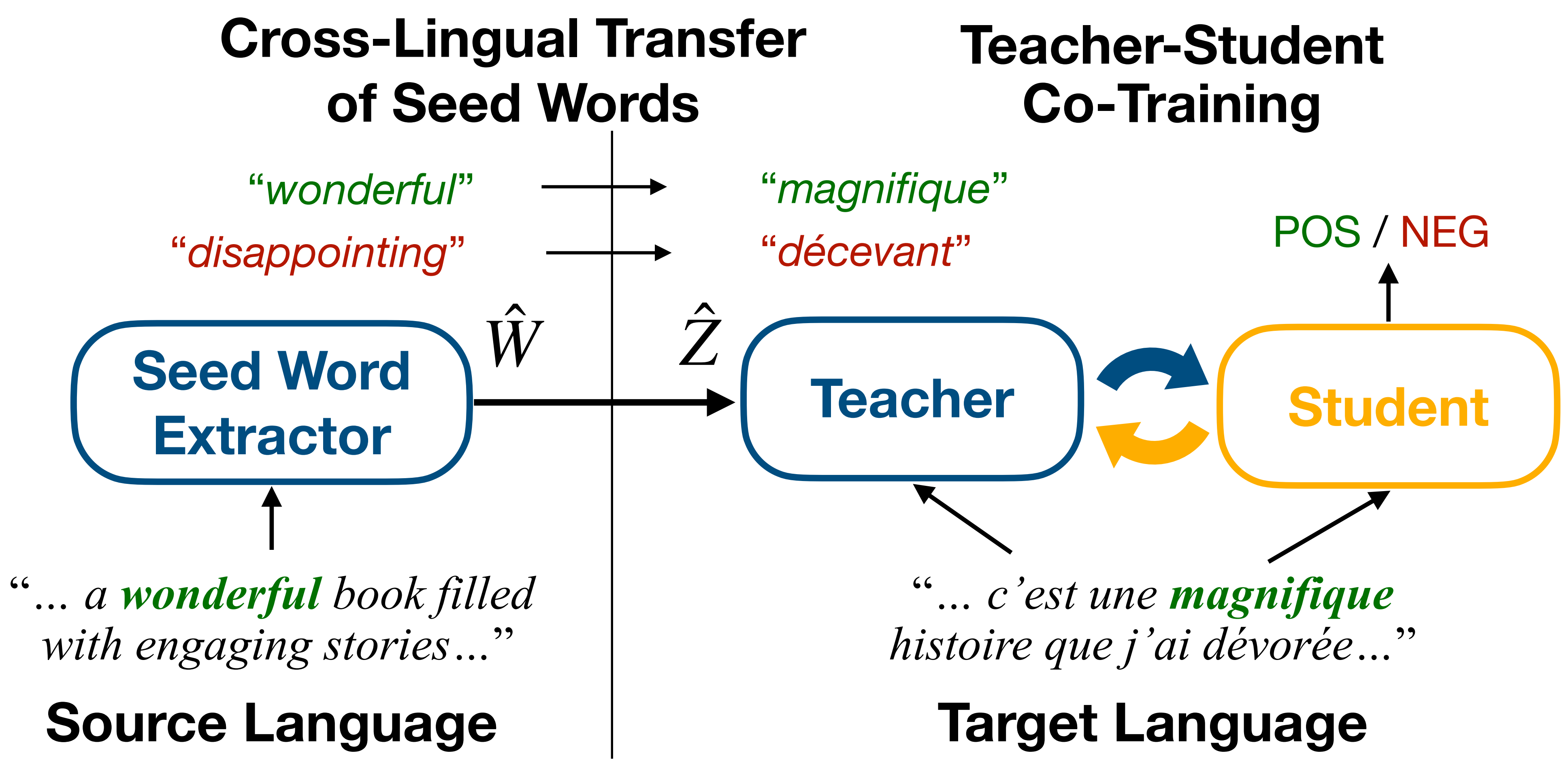
Weakly-Supervised Co-Training in The Target Language

1. Seed-word extraction in the source language
2. Cross-lingual seed weight transfer
3. **Teacher-Student co-training in the target language**
    - Train a more powerful Student on **unlabeled** target documents
    - Student generalizes better than the Teacher



Cross-Lingual Teacher-Student (CLTS)

1. Extract B seed words (non-zero columns in sparse $\hat{W}$)
2. Translate seed words and transfer $\hat{W}$ to $\hat{Z}$
3. Use $\hat{Z}$ as Teacher to (iteratively) train Student



Outline

1. Intro: Cross-Lingual Text Classification

2. Our Approach: Cross-Lingual Teacher-Student (CLTS)

3. Experiments in 18 languages

4. Conclusions

Experiments

18 languages

1. Bulgarian (Bg)
2. German (De)
3. Spanish (Es)
4. Persian (Fa)
5. French (Fr)
6. Croatian (Hr)
7. Hungarian (Hu)
8. Italian (It)
9. Japanese (Ja)
10. Polish (Pl)
11. Portuguese (Pt)
12. Russian (Ru)
13. Sinhalese (Si)
14. Slovak (Sk)
15. Slovenian (Sl)
16. Swedish (Sv)
17. Uyghur (Ug)
18. Chinese (Zh)

4 tasks

1. Topic classification of news documents (MLDoc)

- 4 classes: Corporate/Economics/Government/Markets
- 7 languages: De, Es, Fr, It, Ja, Ru, Zh

2. Sentiment classification of product reviews (CLS)

- 2 classes: positive/negative
- 3 languages: De, Fr, Ja
- 3 product domains per language: books, dvd, music

3. Sentiment classification of tweets (TwitterSent/SentiPers)

- 3 classes: positive/neutral/negative
- 12 languages: Bg, De, Es, Fa, Hr, Hu, Pl, Pt, Sk, Sl, Sv, Ug

4. Medical emergency situation detection (LDC LORELEI)

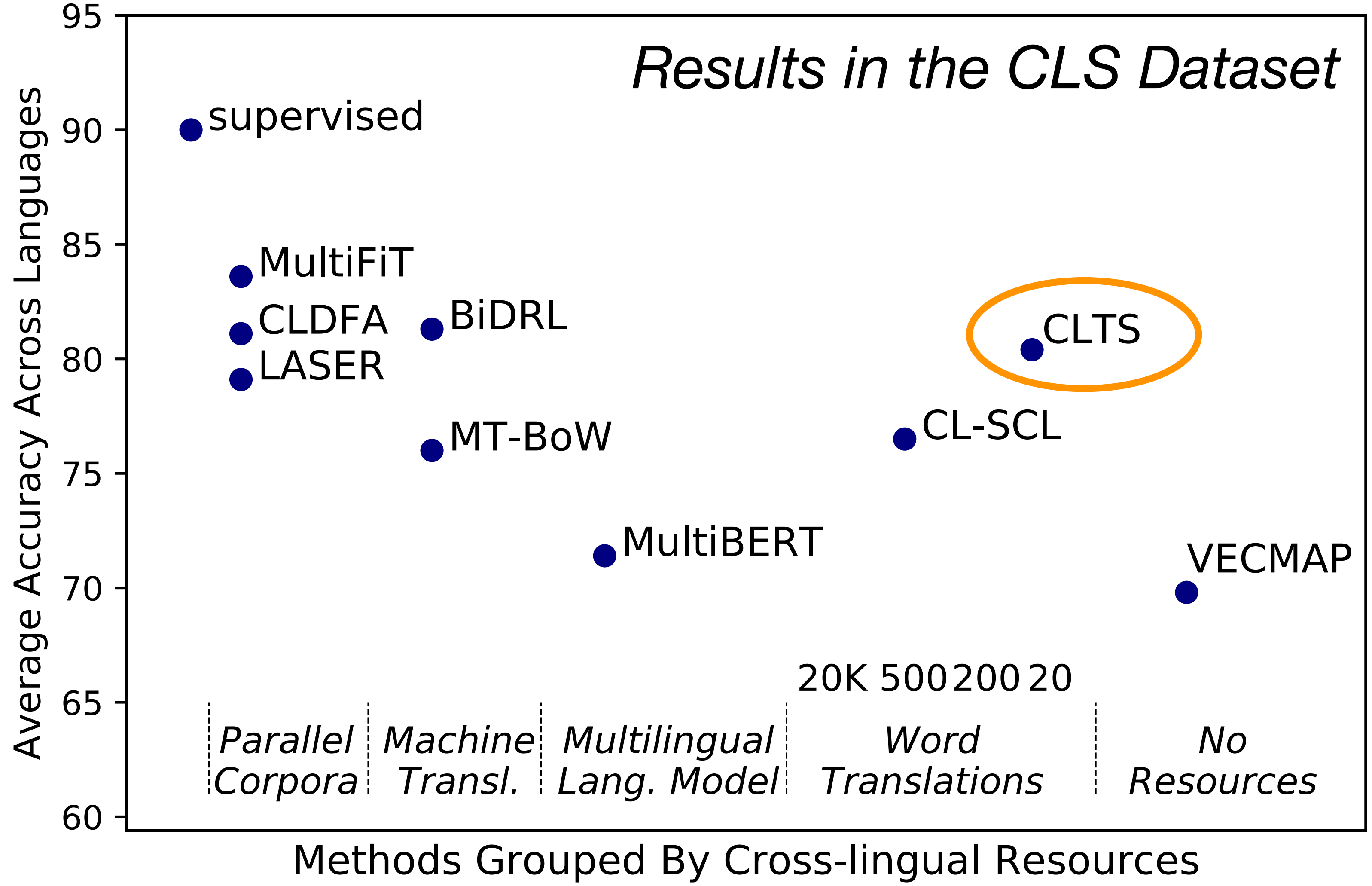
- 2 classes: medical / non-medical
- 2 languages: Si, Ug

Results Summary

- Student outperforms Teacher by 56% (!!!) on average across 18 languages

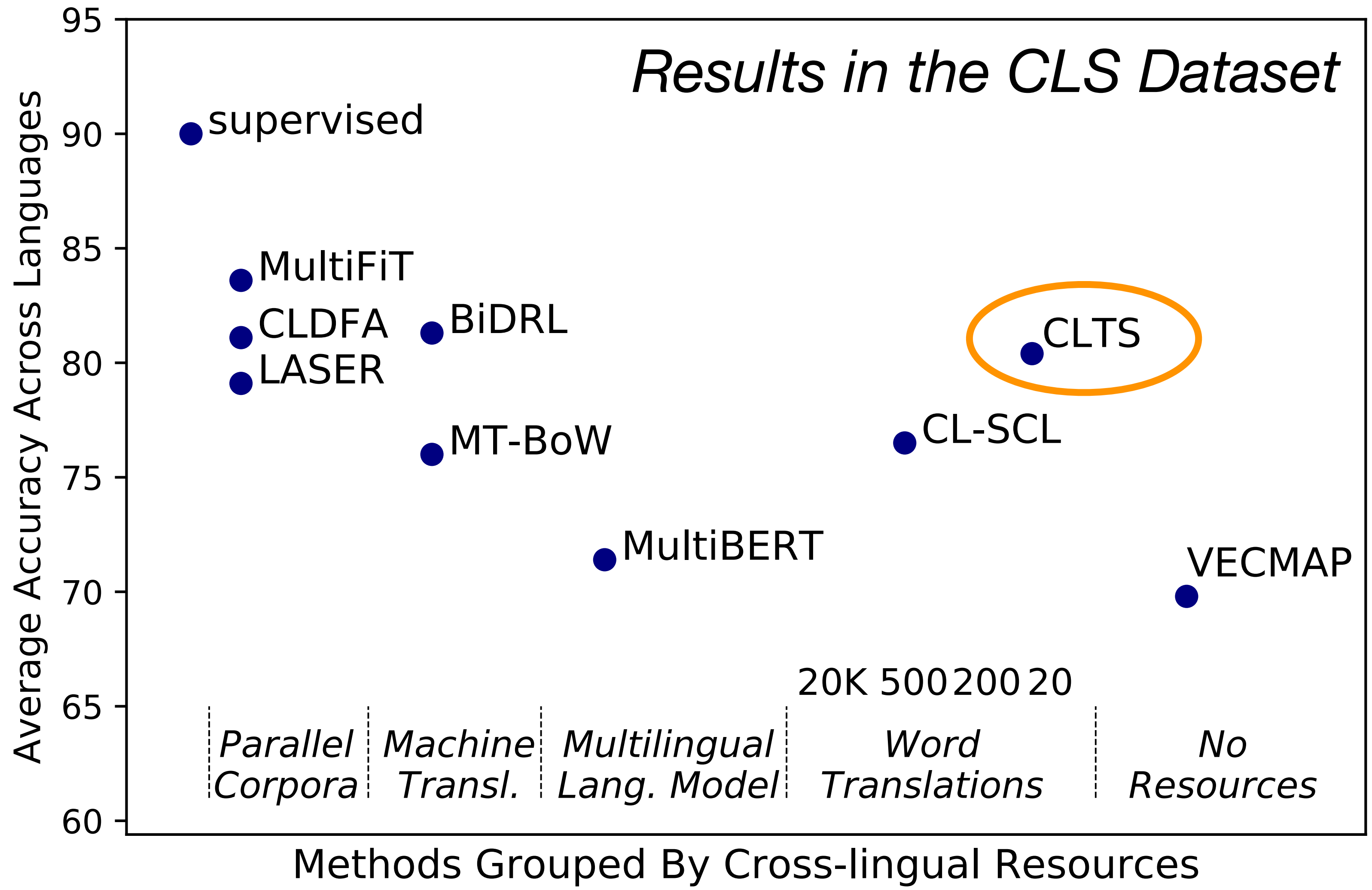
Results Summary

- Student outperforms Teacher by 56% (!!!) on average across 18 languages
- CLTS is effective with as few as 20 word translations



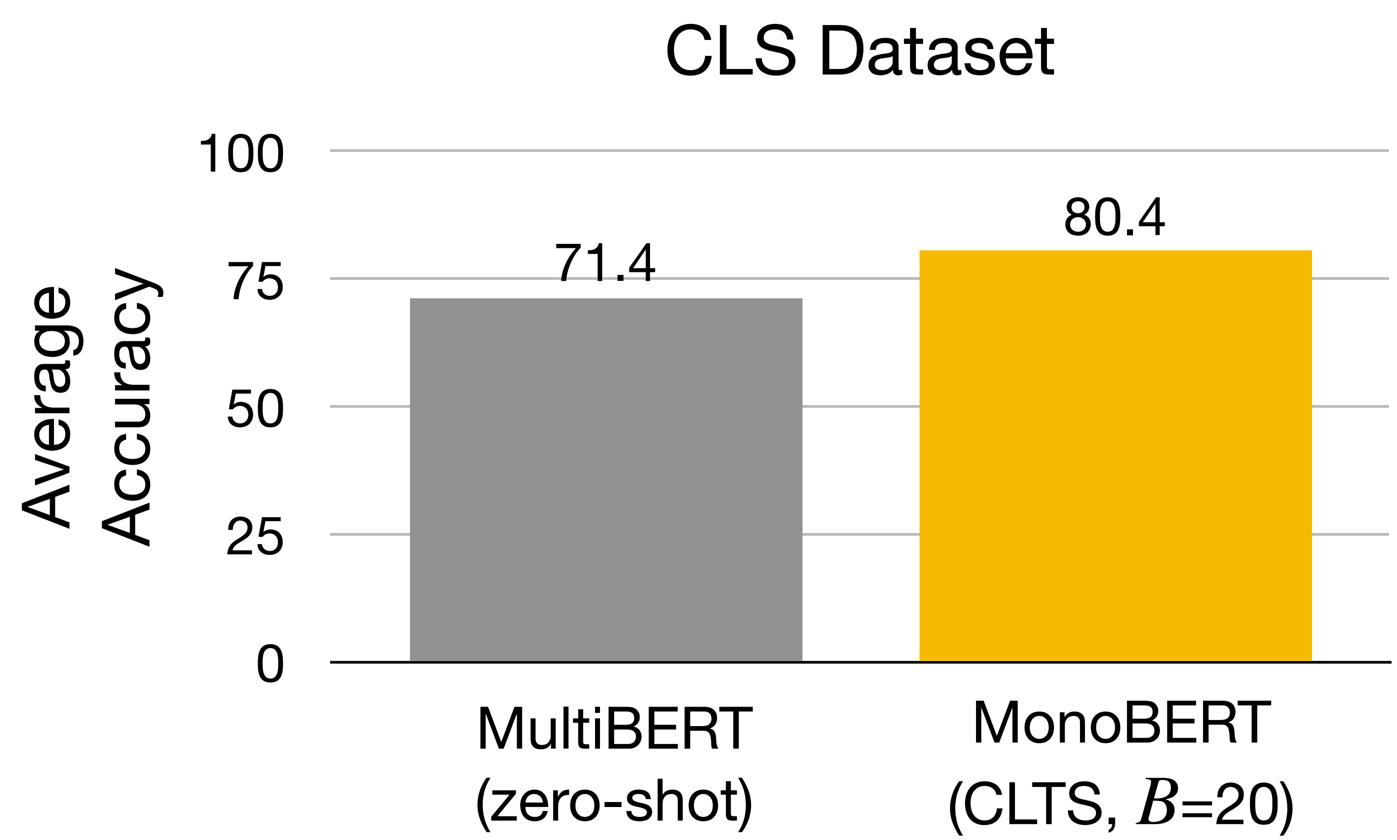
Results Summary

- Student outperforms Teacher by 56% (!!!) on average across 18 languages
- CLTS is effective with as few as 20 word translations
- CLTS sometimes outperforms **even more expensive** approaches by up to 12%



See our paper for more results and ablation experiments!

Transferring Weak Supervision With CLTS > Zero-Shot

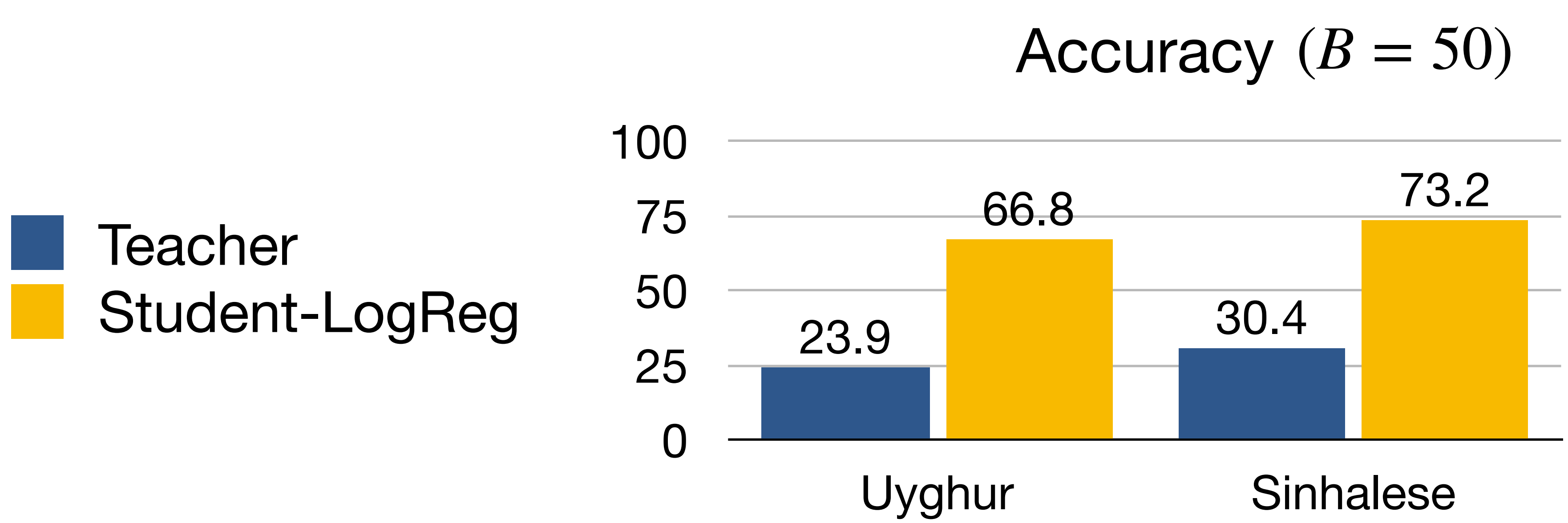


- With just **20 translations** CLTS outperforms zero-shot approaches by 12.6%

Applying CLTS for Low-Resource Languages

- Medical emergency situation detection in Uyghur and Sinhalese

English	->	Uyghur	Sinhalese
1. injured	->	يارىلانغان	තුවාල ලැබුවා
2. attacks	->	ھۇجۇملار	ප්‍රහාර
3. medical	->	medical	වෛද්‍ය
4. crisis	->	كرىزىس	අර්බුදය
5. disease	->	كېسەل	රෝගය
6. malaria	->	بەزگەك كېسەللى	මැලේරියාව
7. health	->	ساغلاملىق	සෞඛ්‍යය
8. injuring	->	يارىلىنىش	තුවාල වීම
9. yemen	->	يەمەن	යේමනය
10. hospitals	->	دوختۇرخانىلار	රෝහල්



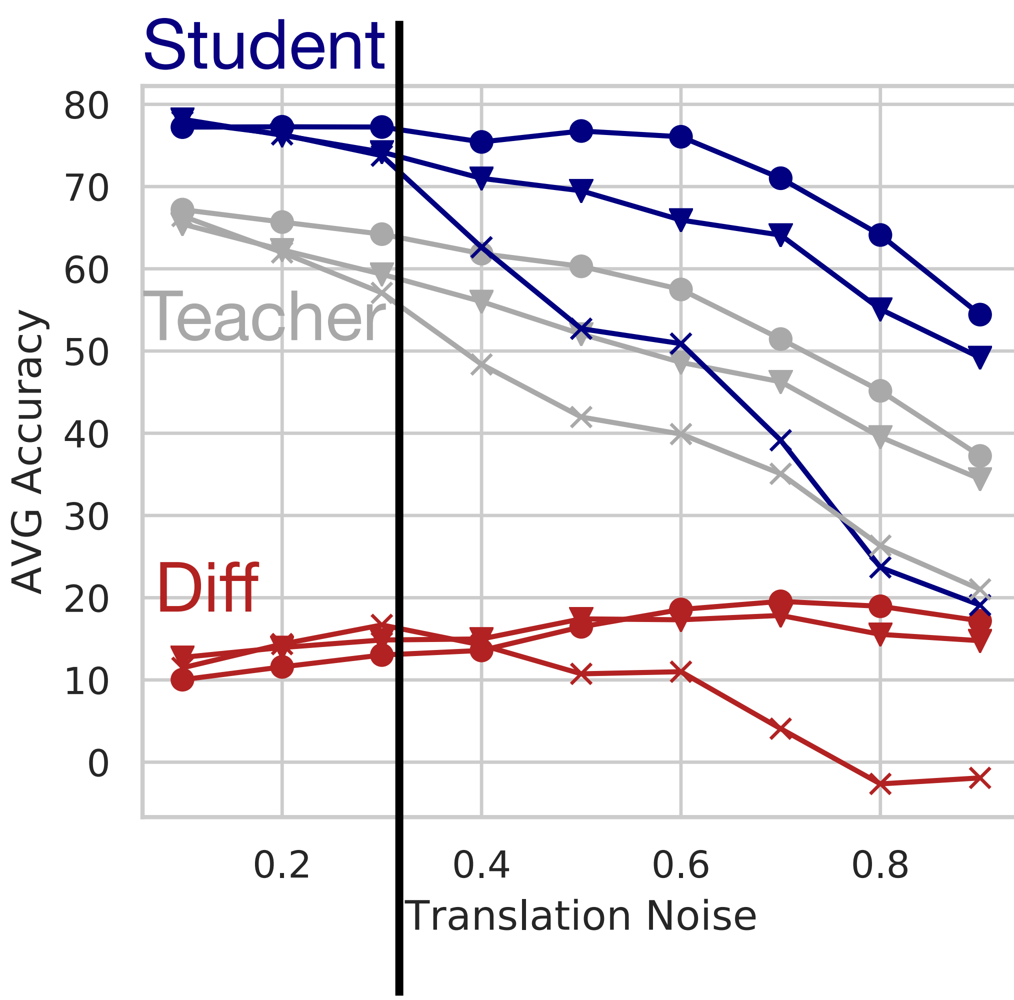
CLTS is Robust To Translation Errors

- Seed words may translate to the wrong words

“Corporate” Topic (En) **“Corporate” Topic (Es)**

“shares” → “comparte” ✗ “acciones” ✓

- Adding simulated translation noise of several types and severity:



- CLTS is effective even with 30% of seed words are translated to wrong words

Outline

1. Intro: Cross-Lingual Text Classification

2. Our Approach: Cross-Lingual Teacher-Student (CLTS)

3. Experiments in 18 languages

4. Conclusions

CLTS Transfers Weak Supervision With Minimal Resources

1. Enable cross-lingual transfer under a **limited translation budget**
    - Use budget as a **sparsity regularizer** when training a source classifier

CLTS Transfers Weak Supervision With Minimal Resources

1. Enable cross-lingual transfer under a **limited translation budget**
    - Use budget as a **sparsity regularizer** when training a source classifier

2. Train **any** target classifier **without labeled target documents**
    - Employ Teacher-Student co-training

CLTS Transfers Weak Supervision With Minimal Resources

1. Enable cross-lingual transfer under a **limited translation budget**
    - Use budget as a **sparsity regularizer** when training a source classifier
2. Train **any** target classifier **without labeled target documents**
    - Employ Teacher-Student co-training
3. Show the benefits of generating weak supervision in **18 languages**
    - CLTS is effective with as few as 20 seed word translations

CLTS Transfers Weak Supervision With Minimal Resources

1. Enable cross-lingual transfer under a **limited translation budget**
    - Use budget as a **sparsity regularizer** when training a source classifier
2. Train **any** target classifier **without labeled target documents**
    - Employ Teacher-Student co-training
3. Show the benefits of generating weak supervision in **18 languages**
    - CLTS is effective with as few as 20 seed word translations

- CLTS can potentially be applied for emerging tasks in low-resource languages

Thank you!

CLTS Code: https://github.com/gkaramanolakis/clts

Contact
gkaraman@cs.columbia.edu
https://gkaramanolakis.github.io

COLUMBIA
UNIVERSITY