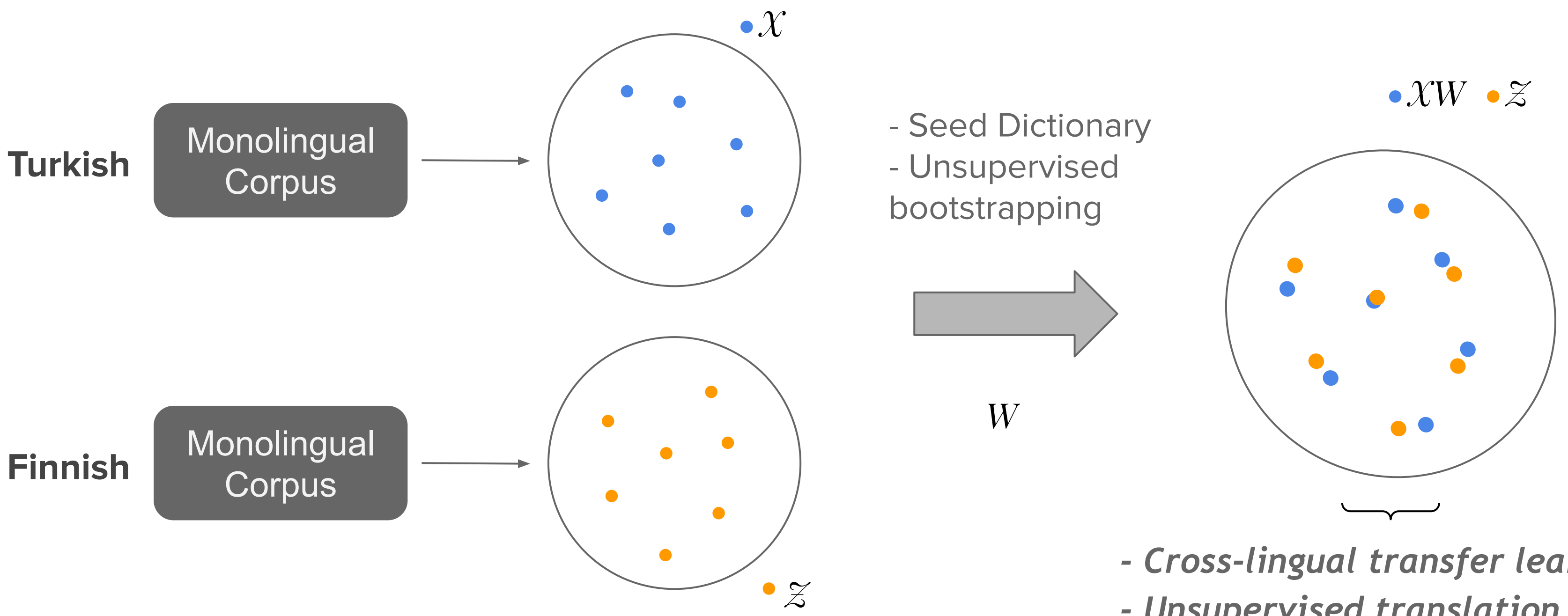
Cross-Lingual Word Embeddings for Morphologically Rich Languages

Ahmet Üstün *Gosse Bouma* *Gertjan van Noord*

University of Groningen

Content

Introduction to CLEs



- *Cross-lingual transfer learning*
- *Unsupervised translation*

Challenges with Rich Morphology

❖ Monolingually, morphological complexity causes high sparsity

Challenges with Rich Morphology

- Monolingually, morphological complexity causes high sparsity
- Cross-Lingually, rich morphology causes inaccurate mappings especially for complex words on morphologically rich language pairs

Challenges with Rich Morphology

- Monolingually, morphological complexity causes high sparsity
- Cross-Lingually, rich morphology causes inaccurate mappings especially for complex words on morphologically rich language pairs
- CLE models are also unable to map an inflected word in the morphologically rich language to a counterpart which corresponds to a phrase in a language with simple morphology.

Challenges with Rich Morphology

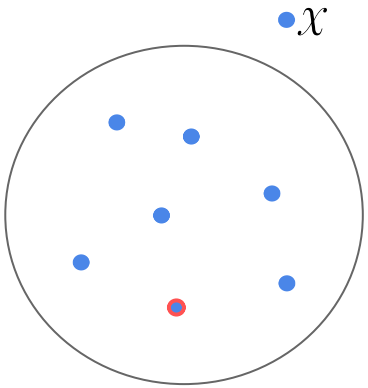
- Monolingually, morphological complexity causes high sparsity
- **Cross-Lingually, rich morphology causes inaccurate mappings especially for complex words on morphologically rich language pairs**
- CLE models are also unable to map an inflected word in the morphologically rich language to a counterpart which corresponds to a phrase in a language with simple morphology.

Motivation to Model

- **Cross-Lingually, rich morphology causes inaccurate mappings especially for complex words on morphologically rich language pairs**

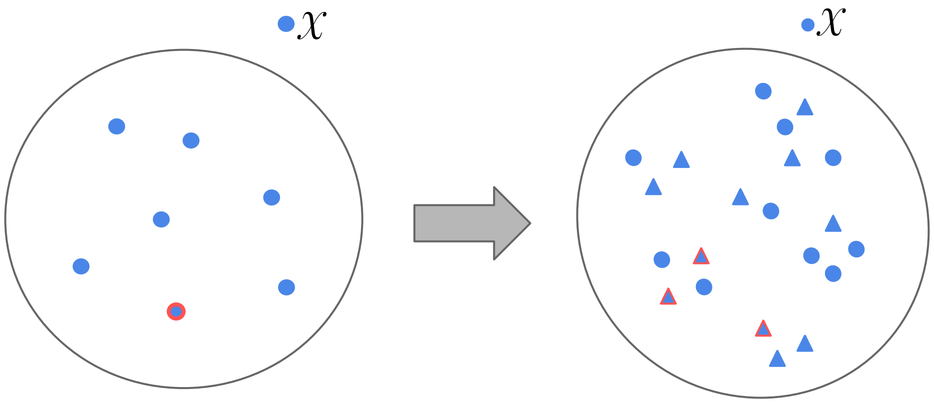
- We propose a morphologically- sensitive cross-lingual word embedding model.
  - Cross-lingual model to learn the morpheme representations in the source languages so that a word can be represented through its morphemes in the target space.
  - Small bilingual dictionary consisting of morphologically complex word pairs.

Morpheme-Based Alignment Model



arabalarda
'in cars'

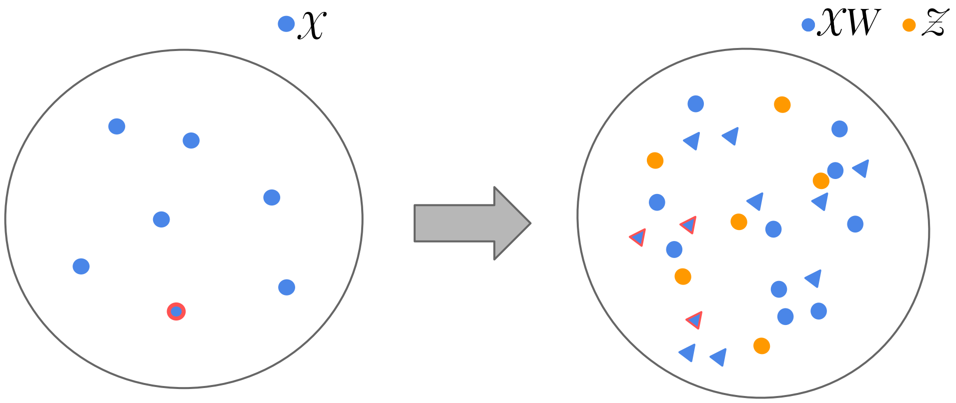
Morpheme-Based Alignment Model



arabalarda
'in cars'

araba - lar - da
car–PLU–LOC
'in cars'

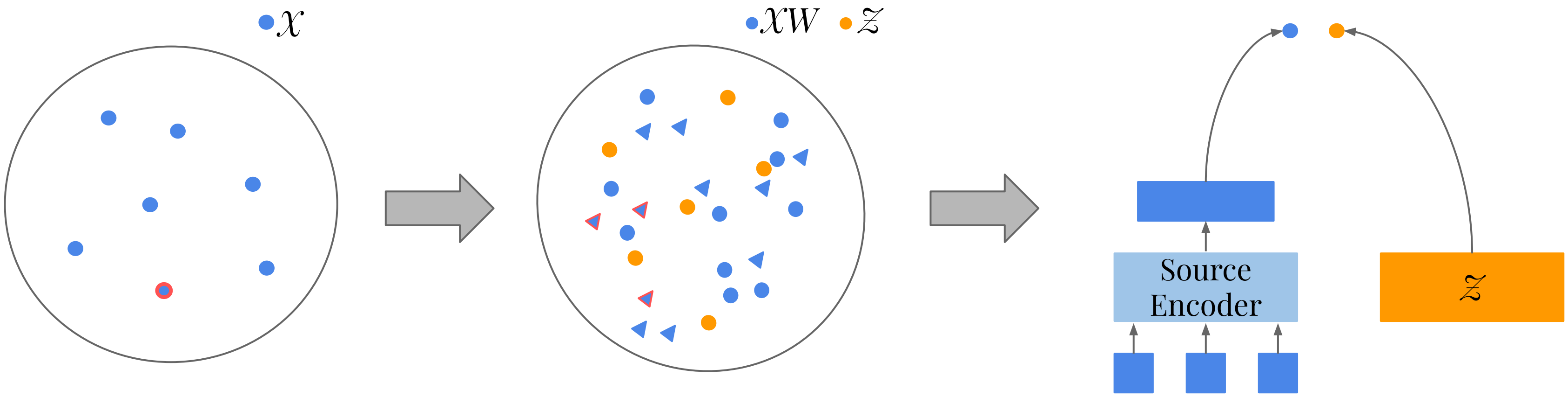
Morpheme-Based Alignment Model



arabalarda
'in cars'

araba - lar - da
car–PLU–LOC
'in cars'

Morpheme-Based Alignment Model



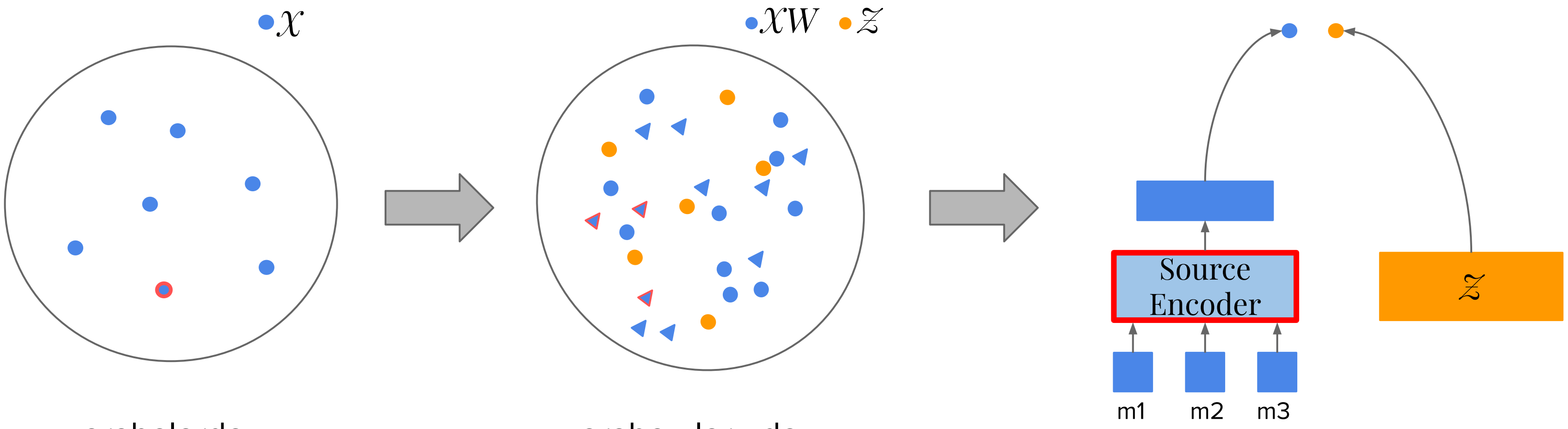
arabalarda
'in cars'

araba - lar - da
car–PLU–LOC
'in cars'

Morphologically sensitive seed dict.

araba - ler - de	autoissa
car–PLU–LOC	car–PLU–LOC

Morpheme-Based Alignment Model



arabalarda
'in cars'

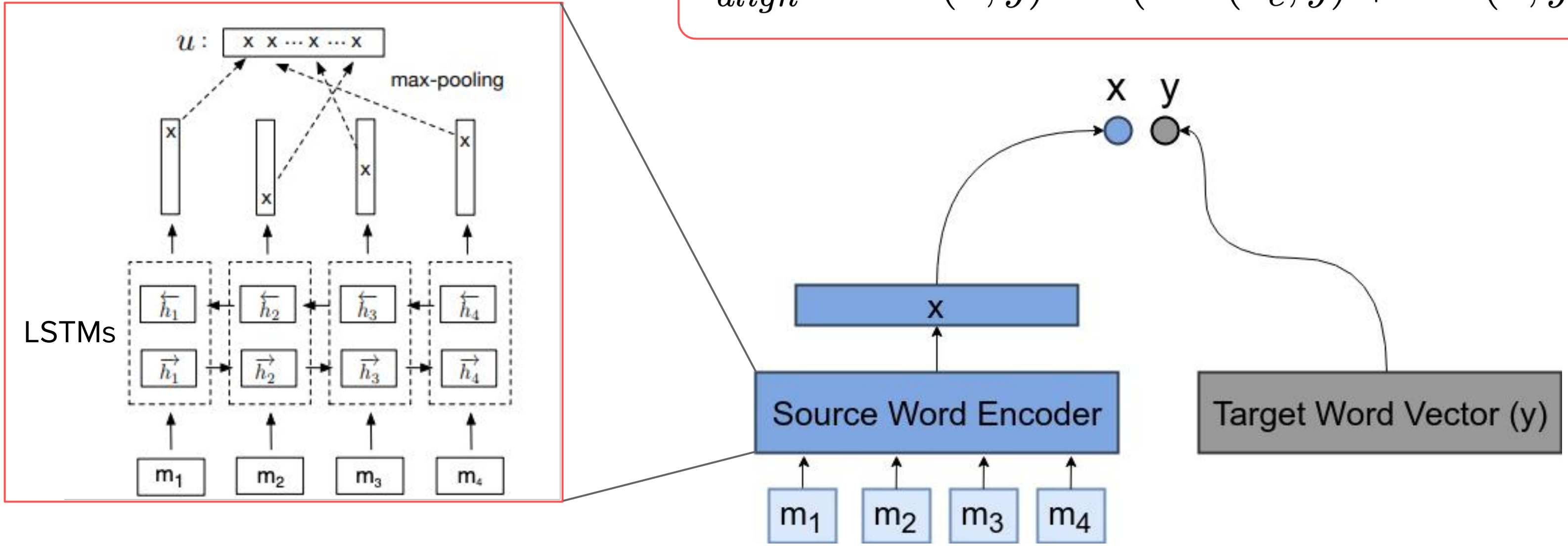
araba - lar - da
car–PLU–LOC
'in cars'

Morphologically sensitive seed dict.

araba - ler - de	autoissa
car–PLU–LOC	car–PLU–LOC

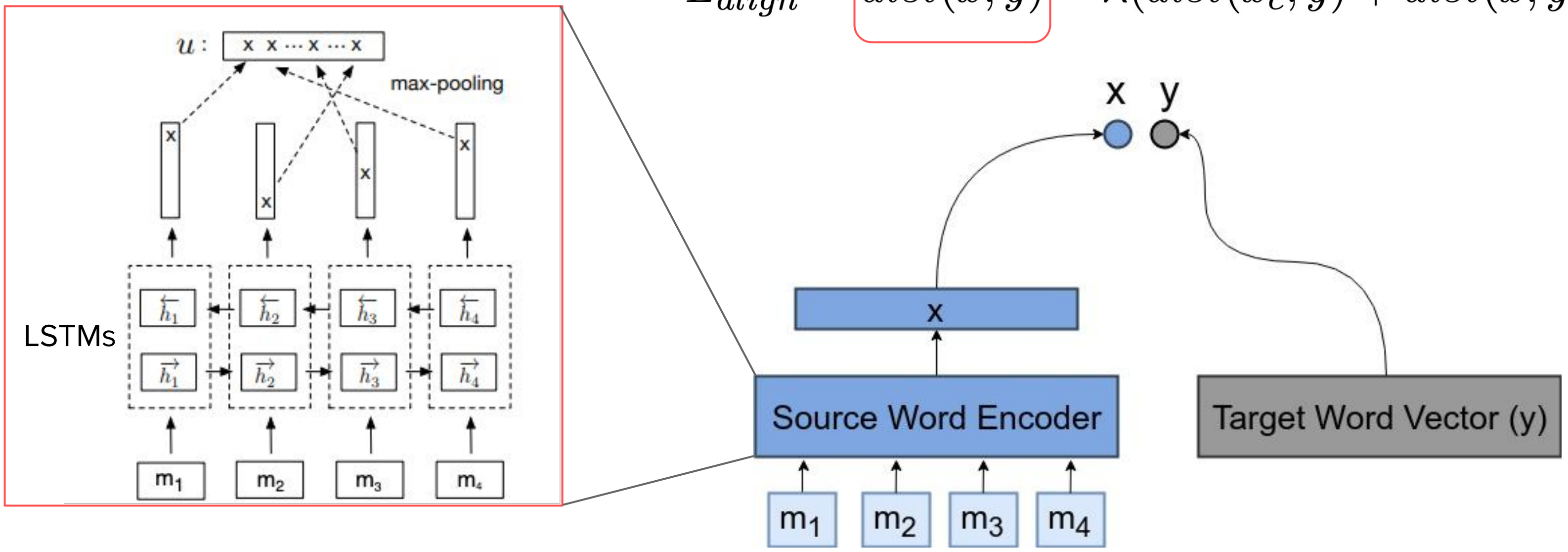
Morpheme-Based Alignment Model

$$L_{align} = dist(x, y) - \lambda(dist(x_c, y) + dist(x, y_c))$$



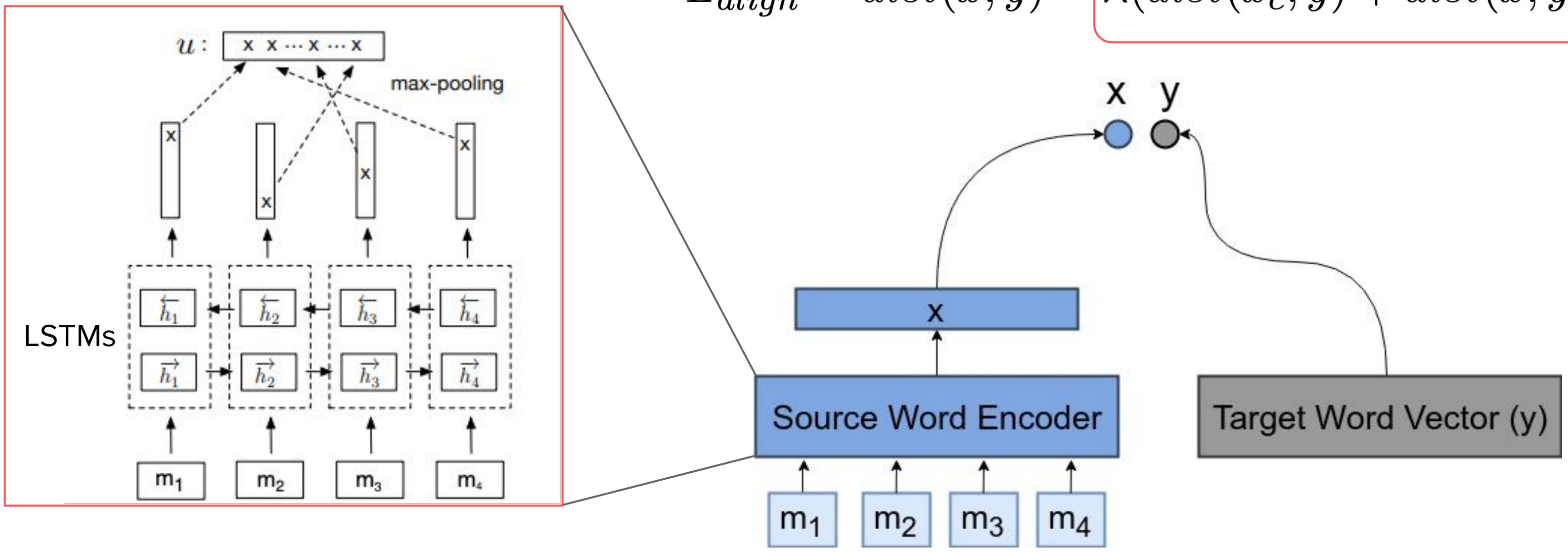
Morpheme-Based Alignment Model

$$L_{align} = \boxed{dist(x,y)} - \lambda(dist(x_c,y) + dist(x,y_c))$$



Morpheme-Based Alignment Model

$$L_{align} = dist(x, y) - \boxed{\lambda(dist(x_c, y) + dist(x, y_c))}$$



Morphologically Sensitive Bilingual Lexicon

- MUSE dataset *(Conneau et al., 2017b)*
- Universal Dependency Treebanks *(Nivre et al., 2016)* and the Universal Morphology *(Sylak-Glassman, 2016)* project

Morphologically Sensitive Bilingual Lexicon

- MUSE dataset *(Conneau et al., 2017b)*
- Universal Dependency Treebanks *(Nivre et al., 2016)* and the Universal Morphology *(Sylak-Glassman, 2016)* project

Turkish	Finnish	
gölgem	varjoni	N;SG;PSS1S
gölgen	varjoasi	N;SG;PSS2S
gölgelerin	varjojen	N;PL;GEN
gölgelerde	varjoissa	N;ESS;PL

Table 1: The inflected wordforms with the same morphological features for the word pair *gölge-varjo* which mean *shadow*

Morphologically Sensitive Bilingual Lexicon

- MUSE dataset *(Conneau et al., 2017b)*
- Universal Dependency Treebanks *(Nivre et al., 2016)* and the Universal Morphology *(Sylak-Glassman, 2016)* project

Attribute	Morphological Classes
Number	Sing, Plu
Polarity	Neg, Pos
Person	{Pss1,Pss2,Pss3}+{Sg,Pl}
Case	{in,on,at}+{Ela,Abl}, Gen, Prt
Tense	Pst, Prs, Imp
Agreement	P1,P2,P3
Voice	Pass
Mood	Ind, Imp, Cond

Table 2: Morphological features which are common in both Turkish and Finnish

Experiments on Bilingual Word Translation

- Procrustes *(Smith et al., 2017; Artetxe et al., 2016)*
- RCSLS - Relaxed cross domain similarity local scaling *(Joulin et al., 2018)*

Model	NN	CSLS
Turkish-Finnish (TR-FI)		
Procrustes	16.54	17.89
RCSLS	18.26	**21.06**
Our model	**20.35**	20.40

Table 3: Bilingual word translation performance of the models at P@1 (%). First three rows show the results after training with Turkish-Finnish morphologically sensitive seed dictionary.

Experiments on Word Similarity *(Monolingual)*

- Morph2Vec *(Üstün et al., 2018)*

  Trained on 200K wiki data

- Fasttext *(Joulin et al., 2017)*

Model	Spearman
Morph2Vec (Üstün et al., 2018)	**52.90**
Our model	42.05
Fasttext (Bojanowski et al., 2017)	20.80

Table 4: The comparison of the Spearman correlation between human judgments and the word similarities obtained by computing the cosine similarity between the learned word embeddings for Turkish.

Analysis

No	Source Word *(Turkish)*	Target Translations *(Finnish)*		
		Procrustes	RCSLS	Our Model
1	öptüm	suutelit *(you kissed)*	suutelen *(I kiss)*	**suutelin *(I kissed)***
2	aileler	perhe *(a family)*	**perheet *(families)***	**perheet *(families)***
3	zamanımız	aikani *(my time)*	aikani *(my time)*	**aikamme *(our time)***
4	acemilerden	aloittelijoilla *(on beginners)*	aloittelijasta *(from beginner)*	**aloittelijoista *(from beginners)***
5	makineler	**koneet *(machines)***	**koneet *(machines)***	koneissa *(in machines)*
6	saatlerde	kellot *(clocks)*	**kelloissa *(in clocks)***	ajoissa *(in times)*

Table 5: Examples comparing the translations of different models which also includes the glosses in English. Bolding indicates the correct translation. In Examples 1-4, our model predicts correct word considering the morphological structure but in the Example 5-6, our model gives wrong translation.

Analysis

No	Source Word *(Turkish)*	Target Translations *(Finnish)*		
		Procrustes	RCSLS	Our Model
1	öptüm	suutelit *(you kissed)*	suutelen *(I kiss)*	**suutelin** ***(I kissed)***
2	aileler	perhe *(a family)*	**perheet** ***(families)***	**perheet** ***(families)***
3	zamanımız	aikani *(my time)*	aikani *(my time)*	**aikamme** ***(our time)***
4	acemilerden	aloittelijoilla *(on beginners)*	aloittelijasta *(from beginner)*	**aloittelijoista** ***(from beginners)***
5	makineler	**koneet** ***(machines)***	**koneet** ***(machines)***	koneissa *(in machines)*
6	saatlerde	kellot *(clocks)*	**kelloissa** ***(in clocks)***	ajoissa *(in times)*

Table 5: Examples comparing the translations of different models which also includes the glosses in English. Bolding indicates the correct translation. In Examples 1-4, our model predicts correct word considering the morphological structure but in the Example 5-6, our model gives wrong translation.

Analysis

No	Source Word *(Turkish)*	Target Translations *(Finnish)*		
		Procrustes	RCSLS	Our Model
1	öptüm	suutelit *(you kissed)*	suutelen *(I kiss)*	**suutelin *(I kissed)***
2	aileler	perhe *(a family)*	**perheet *(families)***	**perheet *(families)***
3	zamanımız	aikani *(my time)*	aikani *(my time)*	**aikamme *(our time)***
4	acemilerden	aloittelijoilla *(on beginners)*	aloittelijasta *(from beginner)*	**aloittelijoista *(from beginners)***
5	makineler	**koneet *(machines)***	**koneet *(machines)***	koneissa *(in machines)*
6	saatlerde	kellot *(clocks)*	**kelloissa *(in clocks)***	ajoissa *(in times)*

Table 5: Examples comparing the translations of different models which also includes the glosses in English. Bolding indicates the correct translation. In Examples 1-4, our model predicts correct word considering the morphological structure but in the Example 5-6, our model gives wrong translation.

Conclusion

- In this work, we extend the simple projection-based cross-lingual embedding (CLE) model to learn a morphology-sensitive transformation between embedding spaces for morphologically rich language pairs.
- We evaluated our model on the bilingual word translation task and compare our results with baselines. Results show that our model learns better alignments for complex word pairs for languages having rich morphology compared to the baseline models.

Thank you !!!

university of groningen

Experiments on Bilingual Word Translation

Model	NN	CSLS
Turkish-Finnish (TR-FI)		
Procrustes	16.54	17.89
RCSLS	18.26	**21.06**
Our model	**20.35**	20.40
TR-FI on English		
Procrustes	12.72	14.89
RCSLS	15.10	17.05

Table 3: Bilingual word translation performance of the models at P@1 (%). First three rows show the results after training with Turkish-Finnish morphologically sensitive seed dictionary. The last two rows present the results when English is used as a pivot language.