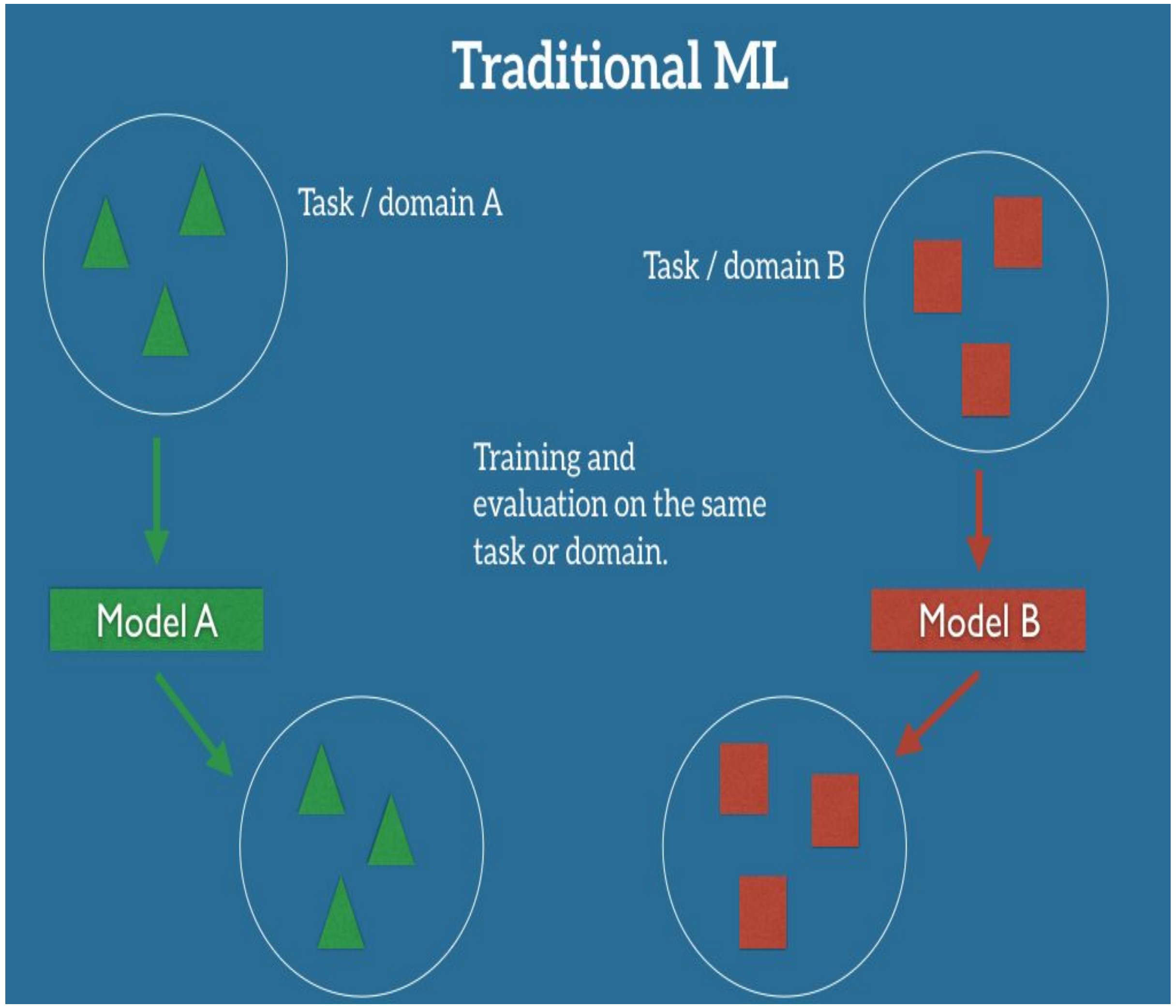
Multilingual Representation for Cross Language NLP

Mentors: Pallika Kanani, Michael Wick

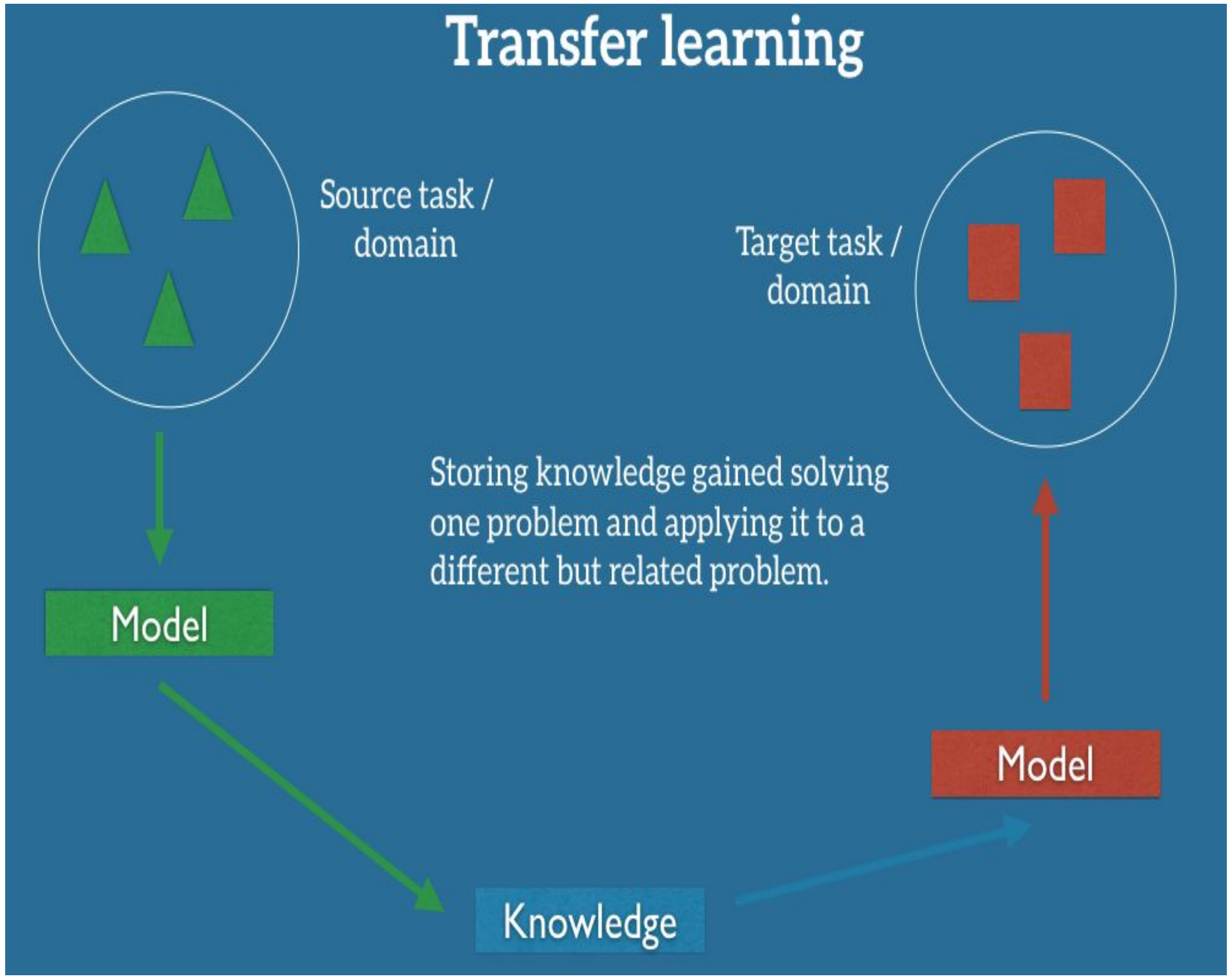
Members: Daniel Thiyagu, Shamya Karumbaiah, Nitin Kishore

Motivation



Traditional ML

Training and evaluation on the same task or domain.



Transfer learning

Storing knowledge gained solving one problem and applying it to a different but related problem.

Introduction

- Motivation - Unavailability of training data in all languages for cross-language NLP
- Goal - Train multilingual word embeddings usable for NLP tasks without retraining in each new language
- Problem - Generalize Multi lingual Word Embeddings and target various NLP Problems like NER, Sentiment
- Approach - Artificial Code Switching

What is ACS ?

Artificial Code Switching

Example:

I have a blue car

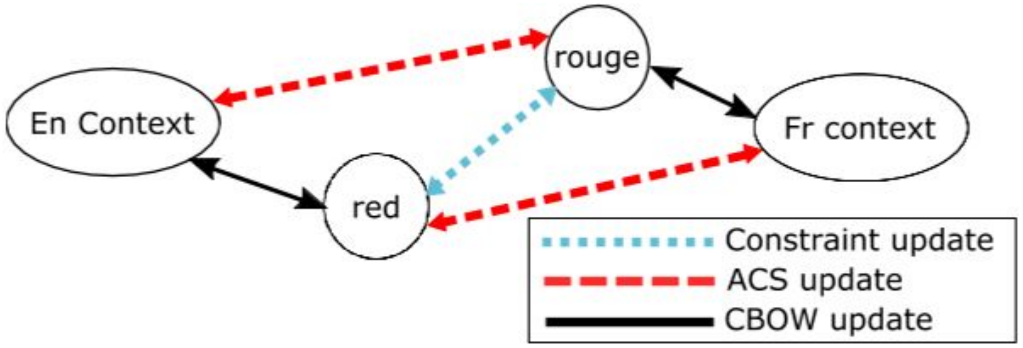
I have un blue car

I have un blue gunda

I have a bleu car

Why is ACS useful?

Distributional Semantics



Similar Word Vectors:

a, un

blue, bleu

car,gunda

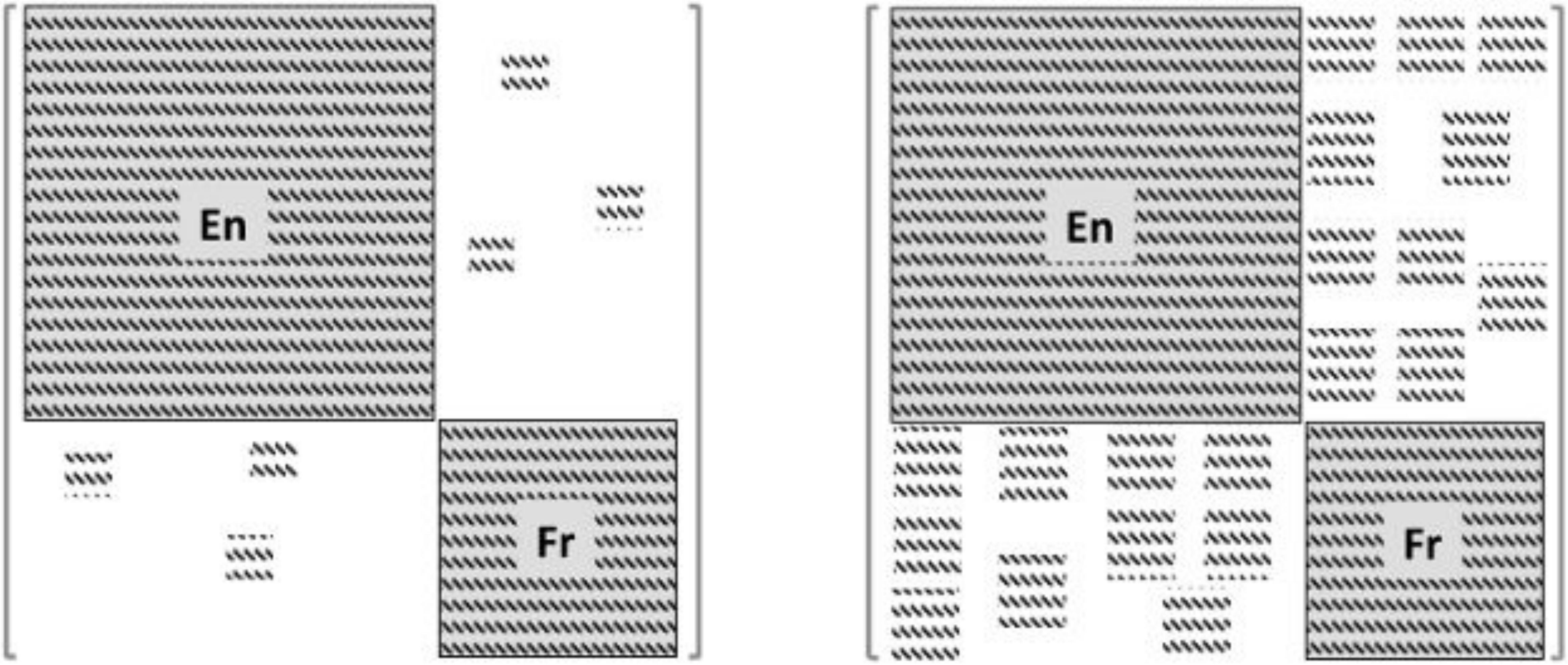
Expectations of ACS

vec("king") - vec("man")+ vec("woman") ≈ vec("queen")

vec("roi") - vec("homme")+ vec("femme") ≈ vec("reine")

vec("roi") - vec("hombre")+ vec("woman") ≈ vec("reine")

Why not just translate entire corpus ?



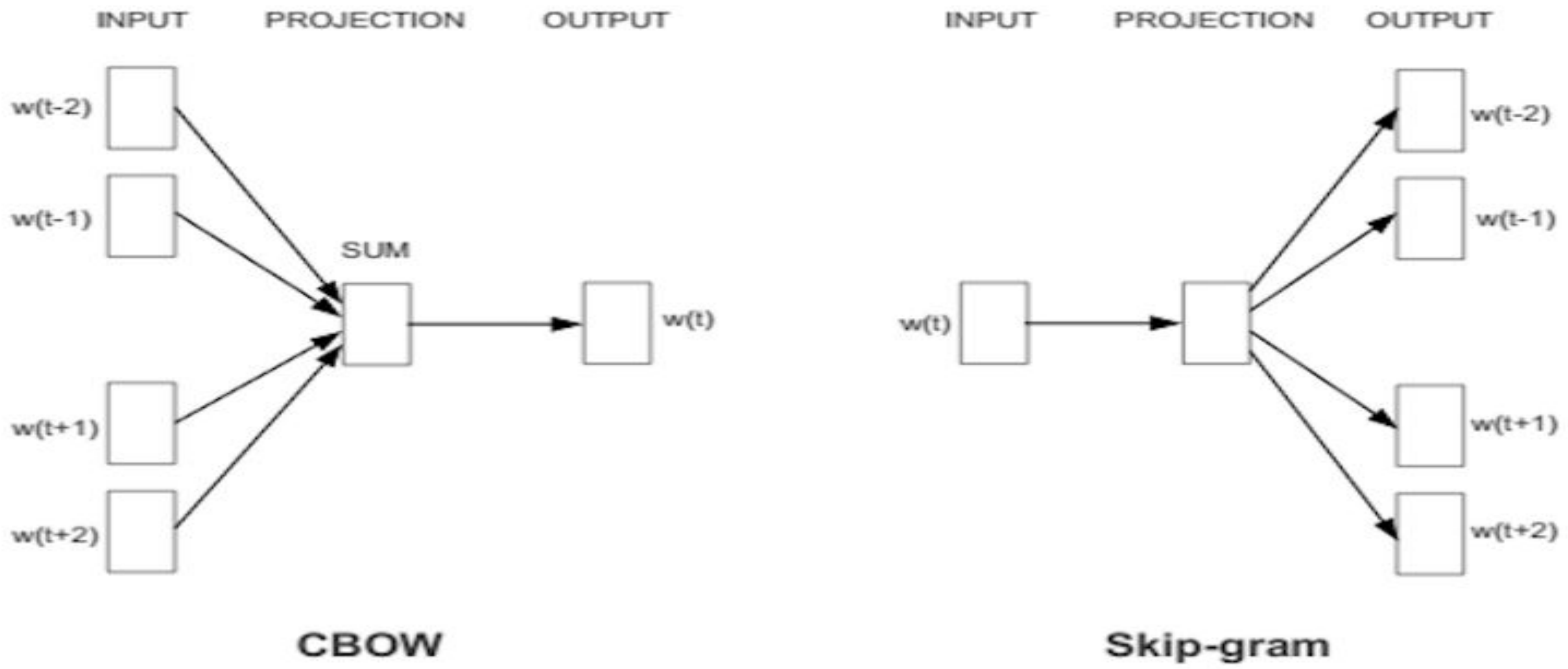
(a) Bilingal co-oc matrix

(b) ACS co-oc matrix.

Word Embeddings ?

CBOW : Works well on Syntactic

Skip Gram : Works well on Semantic



Statistics

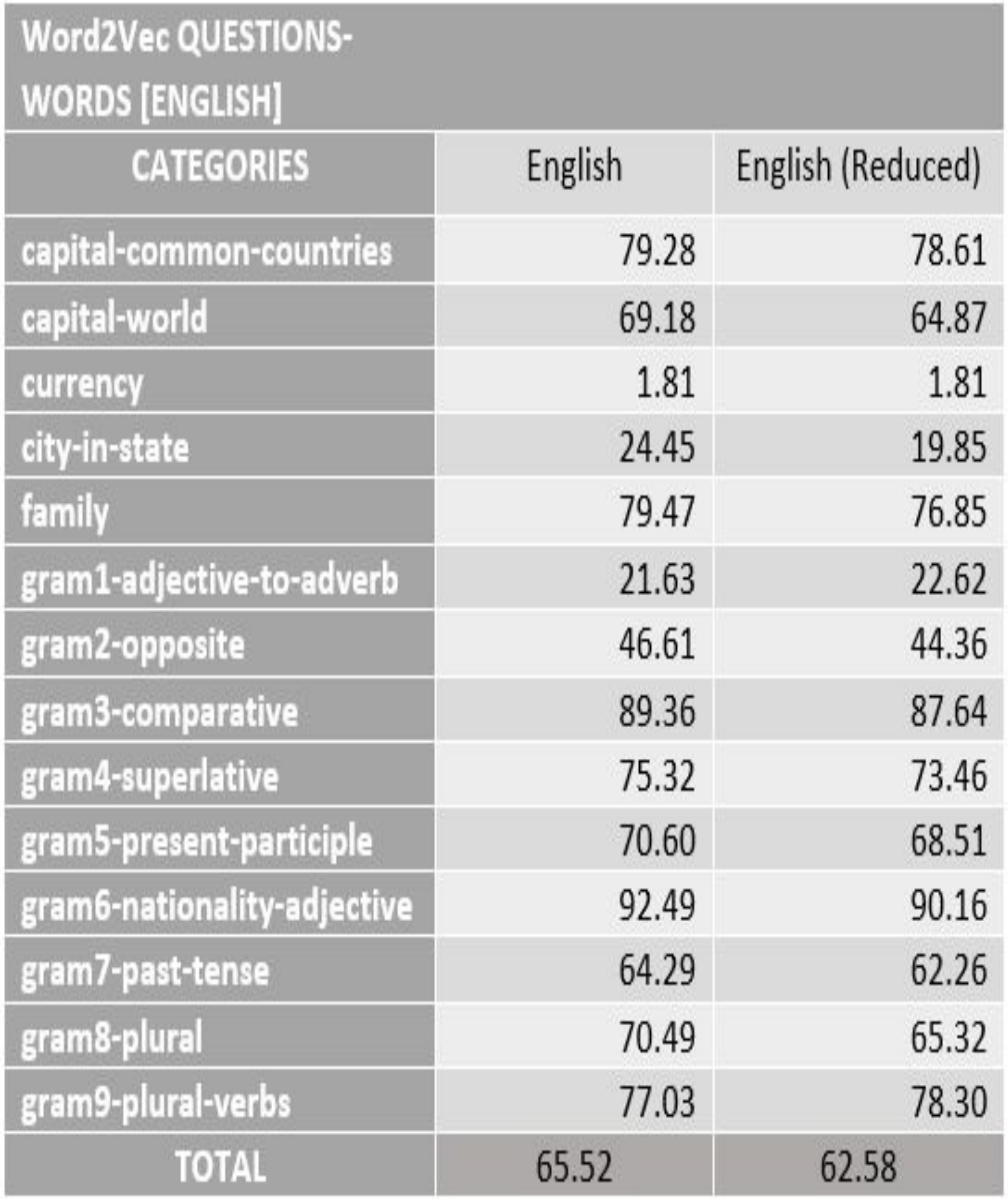
Training multilingual corpus : CBOW

Languages : French, Italian, English, Spanish

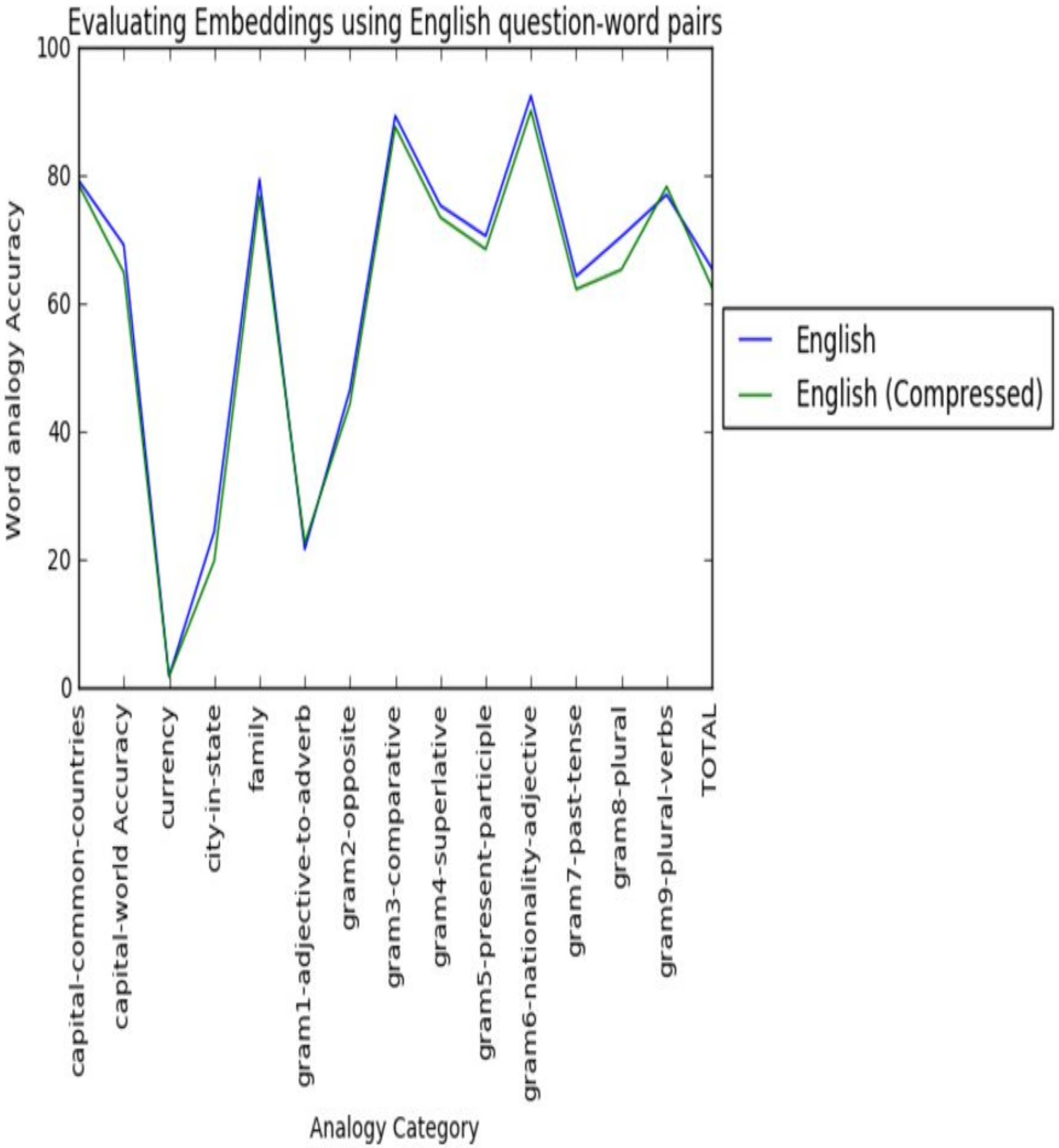
DataSet size: 9GB of Articles on Wikipedia

Factorie Toolkit for training word embeddings

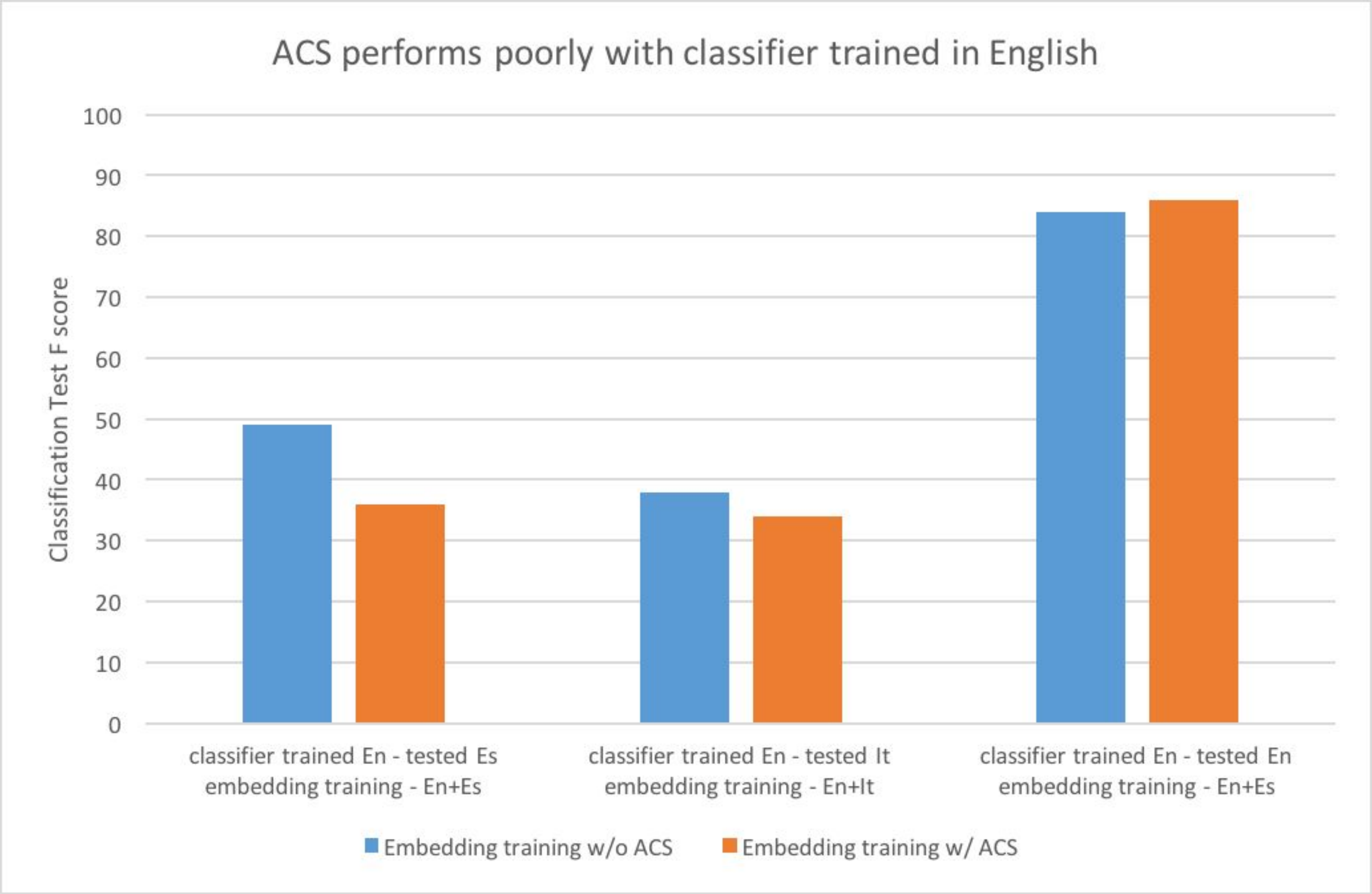
Intrinsic Evaluation



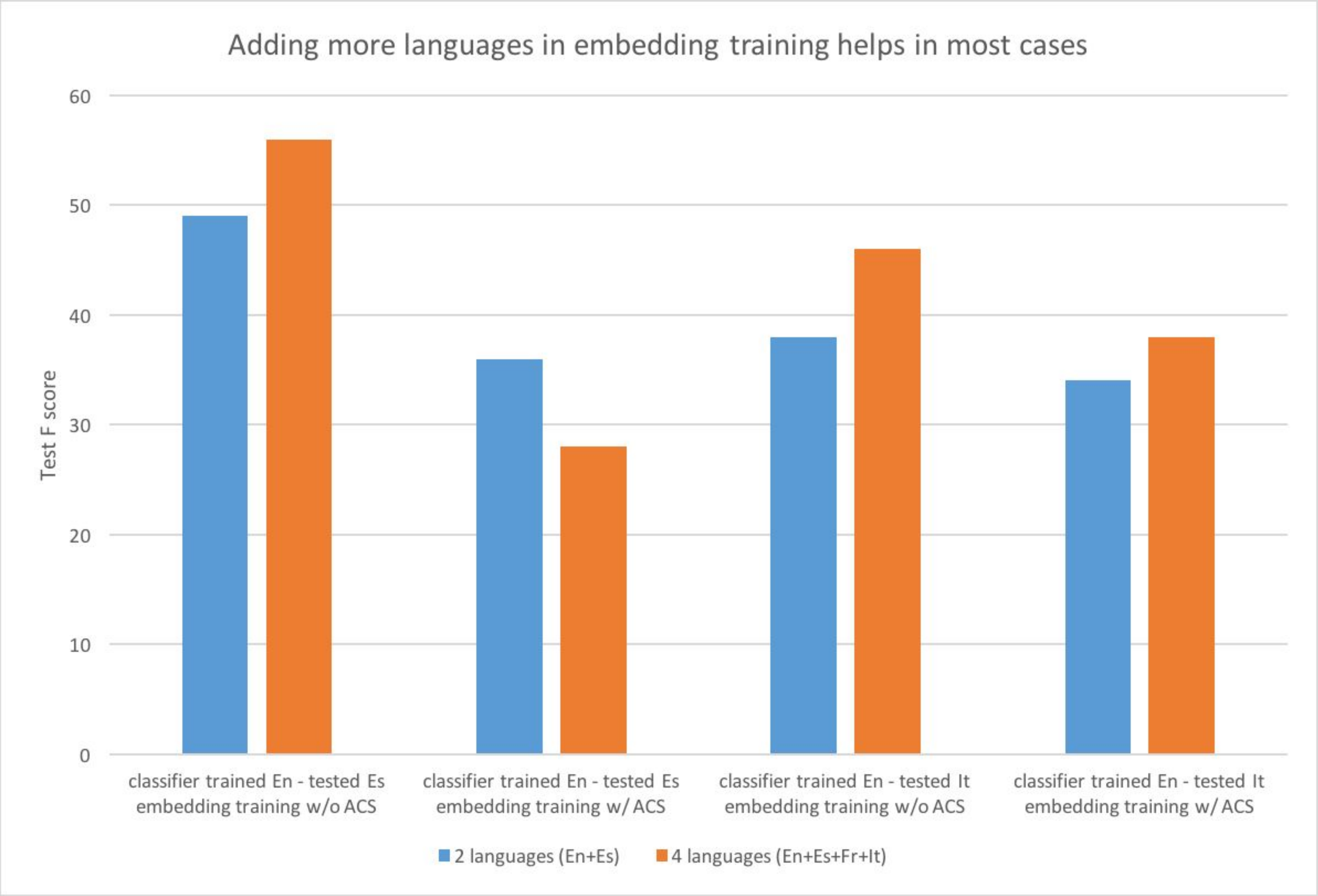
Word2Vec QUESTIONS-WORDS [ENGLISH]		
CATEGORIES	English	English (Reduced)
capital-common-countries	79.28	78.61
capital-world	69.18	64.87
currency	1.81	1.81
city-in-state	24.45	19.85
family	79.47	76.85
gram1-adjective-to-adverb	21.63	22.62
gram2-opposite	46.61	44.36
gram3-comparative	89.36	87.64
gram4-superlative	75.32	73.46
gram5-present-participle	70.60	68.51
gram6-nationality-adjective	92.49	90.16
gram7-past-tense	64.29	62.26
gram8-plural	70.49	65.32
gram9-plural-verbs	77.03	78.30
TOTAL	65.52	62.58



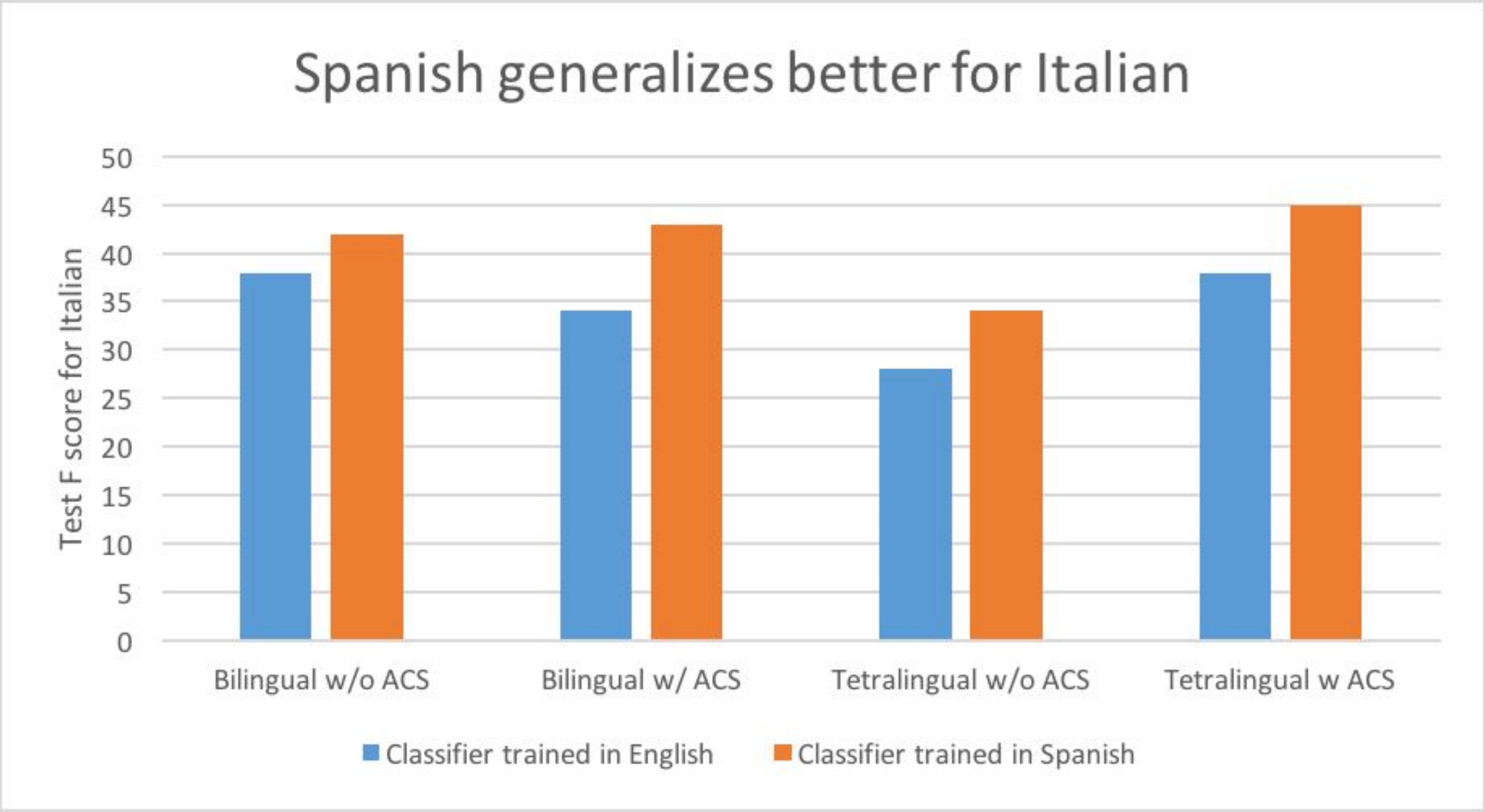
Extrinsic Evaluation



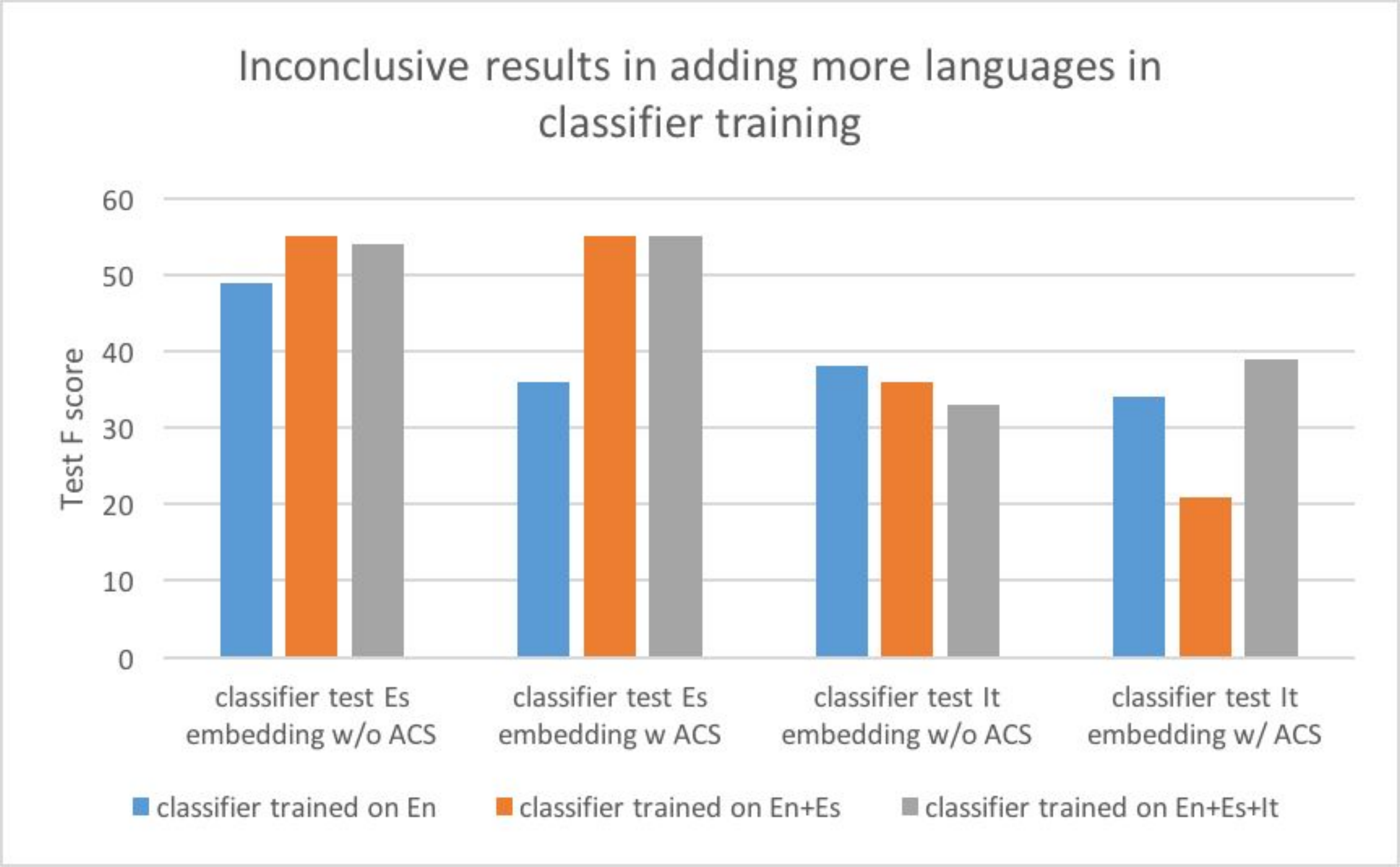
Extrinsic Evaluation



Extrinsic Evaluation



Extrinsic Evaluation



Conclusion

- Certain pairs of languages are better at generalizing (Spanish - Italian)

- ACS performs poorly - multiple sweeps of code switching needed

- Adding more languages for embedding training generally seems to help

- Adding more languages to classifier training has mixed results

Future Work

- Improve ACS (word sense disambiguation, evaluation with concept cosine similarity, phrase level switch)

- ACS hyper parameter sweep (switch threshold)

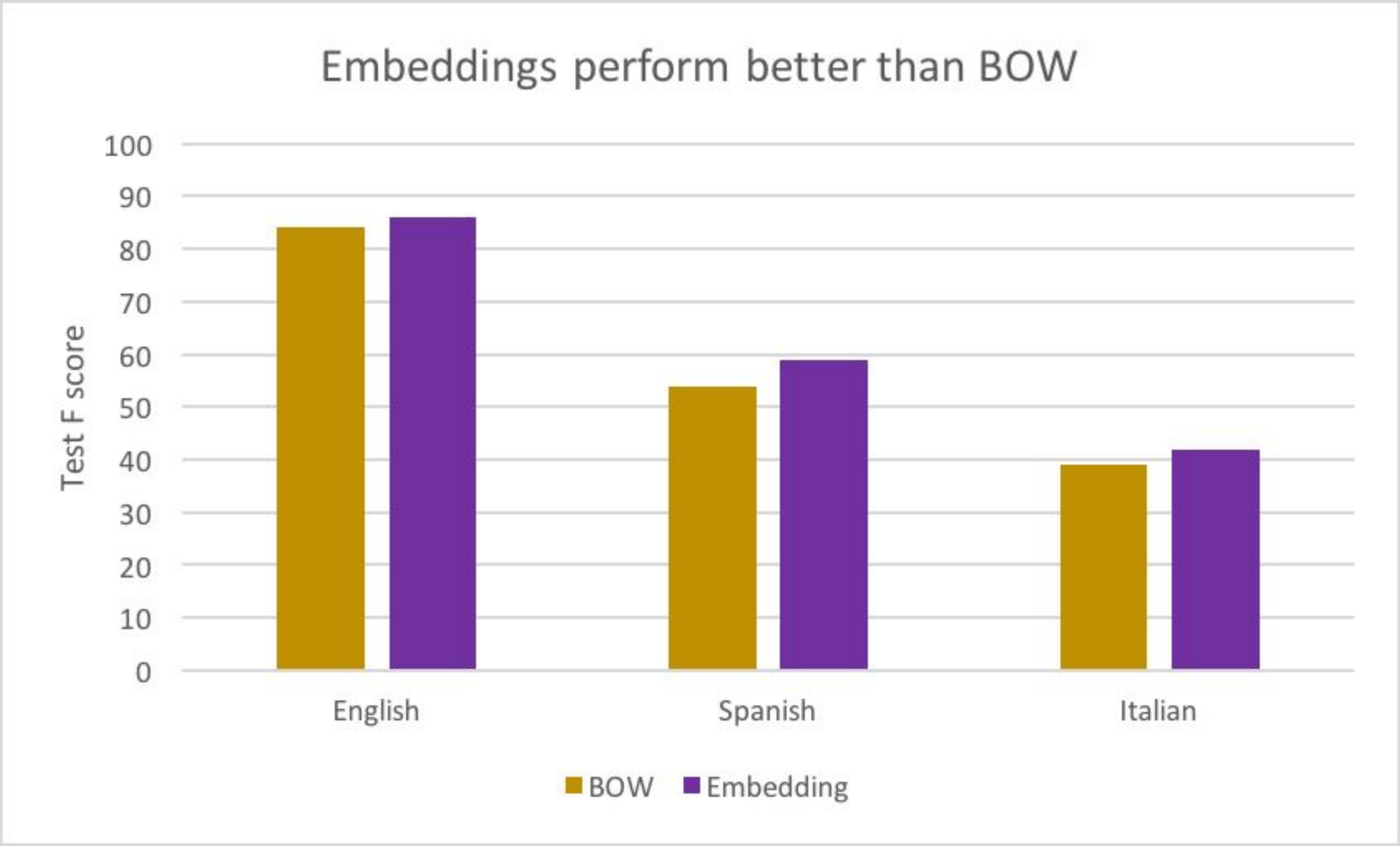
- Test performance and generalizability (evaluation on other NLP tasks, new languages)

- Improve evaluation task approach (phrase-level/ document-level features, better algorithms)

Questions ?

Diolch
Kiitos
umesc
Sheun
Shnorhakalutiun
Kasih
Mamnoon
Todah
Dekuju/Dekujeme
Teşekkür
Hvala
Shokriya
Dziękuję
Shokran
Spaas
Mul
Ačí
Gamsahapnida
Ngiyabonga
Cam
Xie
Dank
Takk
or
Gra
Dhanyavaad
Kop
Salamat
Merci
al
Grazie
Faleminderit
Dakujem
Waad
Dhanyavadaalu
Daw
Dankie
Go
Dhanyavad
Khopjai
Kruthagnathalu
ThankYou
kun
Arigatou
krap
Or
Dhonnobaad
Tack
ederim
Asante
Shukriya
Grazzi
raibh
Gracias
Nandree
Blagodariya
Gomapsupnida
Euxaristo
Kun
Hain
Dhan
Fyrir
Terima
Enkosi
Danke
dank
daa

Extrinsic Evaluation


Embeddings perform better than BOW
Test F score
English
Spanish
Italian
BOW
Embedding