Does Syntactic Knowledge Transfer Cross-Lingually in Multilingual Neural Language Models?

Prajit Dhar, **Arianna Bisazza**

Universiteit Leiden

NWO

Introduction

- Neural NLP models can be trained simultaneously on mix of languages

Introduction

- Neural NLP models can be trained simultaneously on mix of languages
  - less models to maintain
  - easy way to transfer knowledge from high- to low-resource languages
  - success story: Google's multilingual NMT (Johnson et al., 2017)
- Multilingual models typically evaluated on downstream tasks

Introduction

- Neural NLP models can be trained simultaneously on mix of languages
  - less models to maintain
  - easy way to transfer knowledge from high- to low-resource languages
  - success story: Google's multilingual NMT (Johnson et al., 2017)
- Multilingual models typically evaluated on downstream tasks

- *What kind of knowledge actually transfers across languages?*
- *What are the conditions for transfer to happen?*

Introduction

- Neural NLP models can be trained simultaneously on mix of languages
  - less models to maintain
  - easy way to transfer knowledge from high- to low-resource languages
  - success story: Google's multilingual NMT (Johnson et al., 2017)
- Multilingual models typically evaluated on downstream tasks

- *What kind of knowledge actually transfers across languages?*
- *What are the conditions for transfer to happen?*

Specifically, we look at **syntactic transfer**...

First Experiment

- Inspiration: cross-linguistic influence in human subjects widely studied
    - e.g. bilingual subjects newly exposed to passive voice in L1 tend to use it in L2 (Hartsuiker et al., 2004)

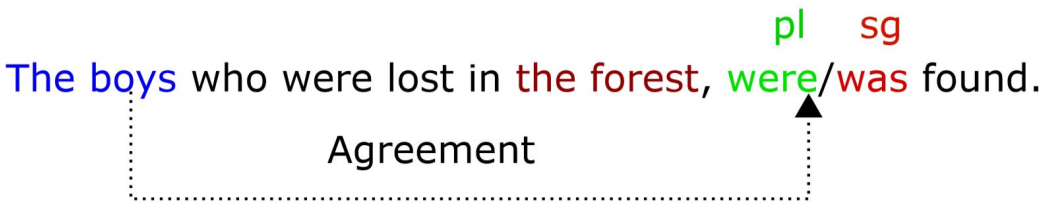
First Experiment

- Inspiration: cross-linguistic influence in human subjects widely studied
    - e.g. bilingual subjects newly exposed to passive voice in L1 tend to use it in L2 (Hartsuiker et al., 2004)
- NLP Task: Long-distance agreement prediction by LM score
- Train LM on mix of 2 languages: (large) helper L1 + (small) target L2

- *Does exposure to L1 improve accuracy on L2?*

About the Task

- Long-range number agreement task (Gulordava et. al., 2018)



- Two conditions:
  - Original: sentences as extracted from corpus
  - Nonsense: "*The trucks who were welcomed in the pocket, were solved.*"
- Available in English, Italian, Hebrew, Russian (non-parallel)

Setup

- Two-Layer LSTM
- Embed/hidden size: 650
- L1: French, L2: Italian/Russian
- Wikipedia corpus
- Note: Training objective is inherently monolingual! Sharing (mostly) takes place at hidden layers

Setup

- Two-Layer LSTM
- Embed/hidden size: 650
- L1: French, L2: Italian/Russian
- Wikipedia corpus
- Note: Training objective is inherently monolingual! Sharing (mostly) takes place at hidden layers

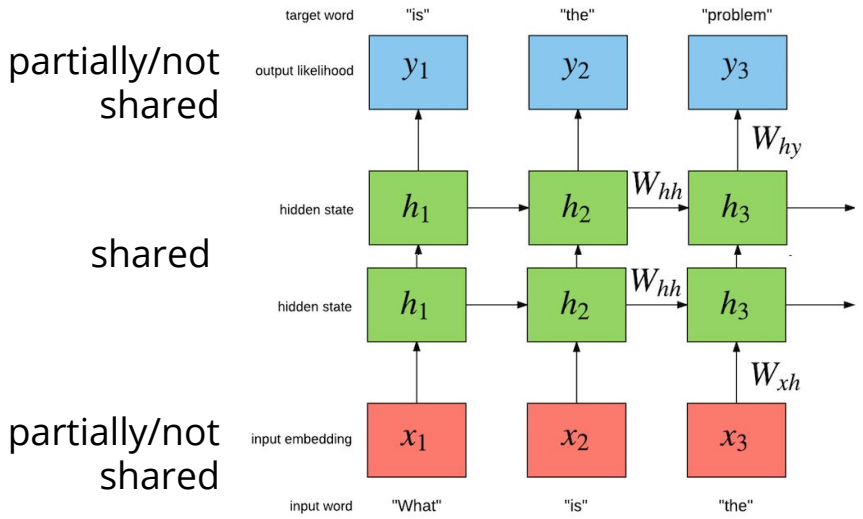
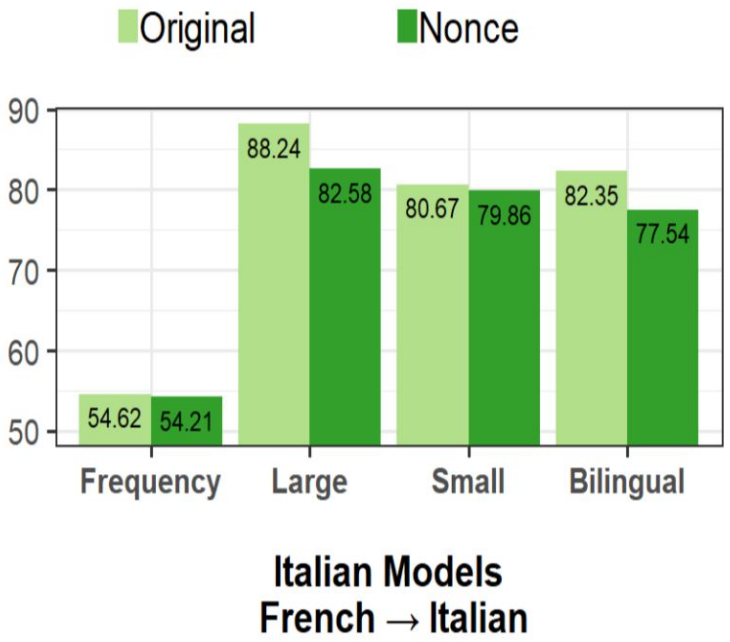
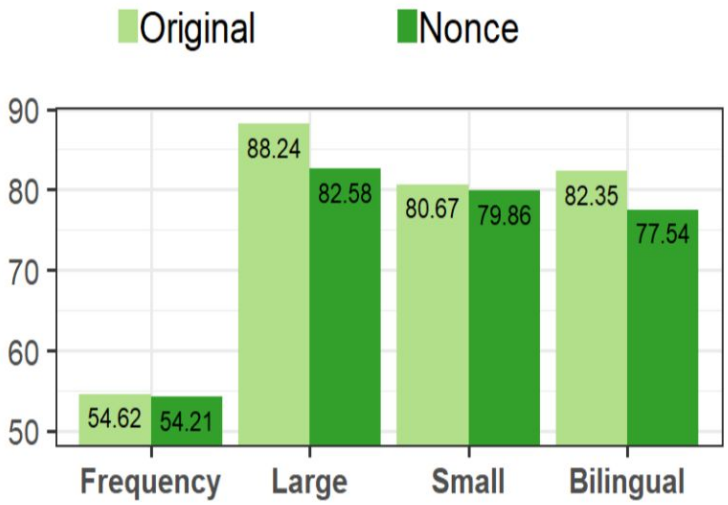


Image: http://torch.ch/blog/2016/07/25/nce.html

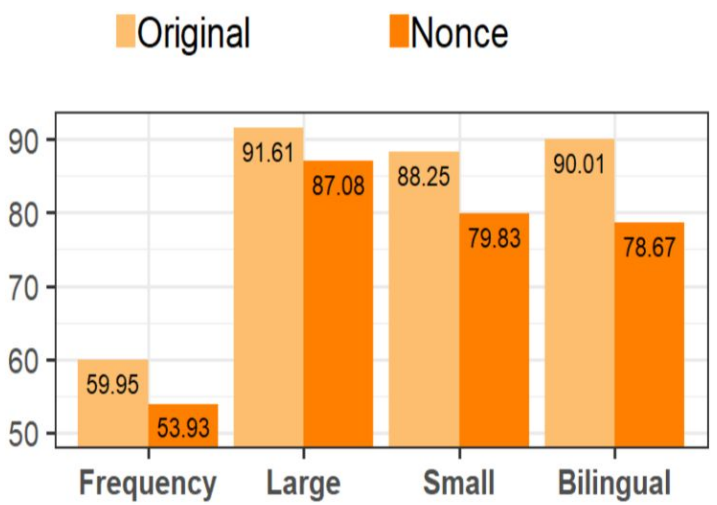
Results


Original
Nonce
90
80
70
60
50
88.24
82.58
80.67
79.86
82.35
77.54
54.62
54.21
Frequency
Large
Small
Bilingual
Italian Models
French → Italian

Results



Italian Models
French → Italian



Russian Models
French → Russian

Consolidating Experiments

Varying Conditions: 1) Vocabulary Overlap

<FR> FR_Amsterdam FR_est FR_la FR_capitale.

<IT> IT_Amsterdam IT_è IT_una IT_città IT_olandese.

No Overlap

<FR> Amsterdam est la capitale.

<IT> Amsterdam è una città olandese.

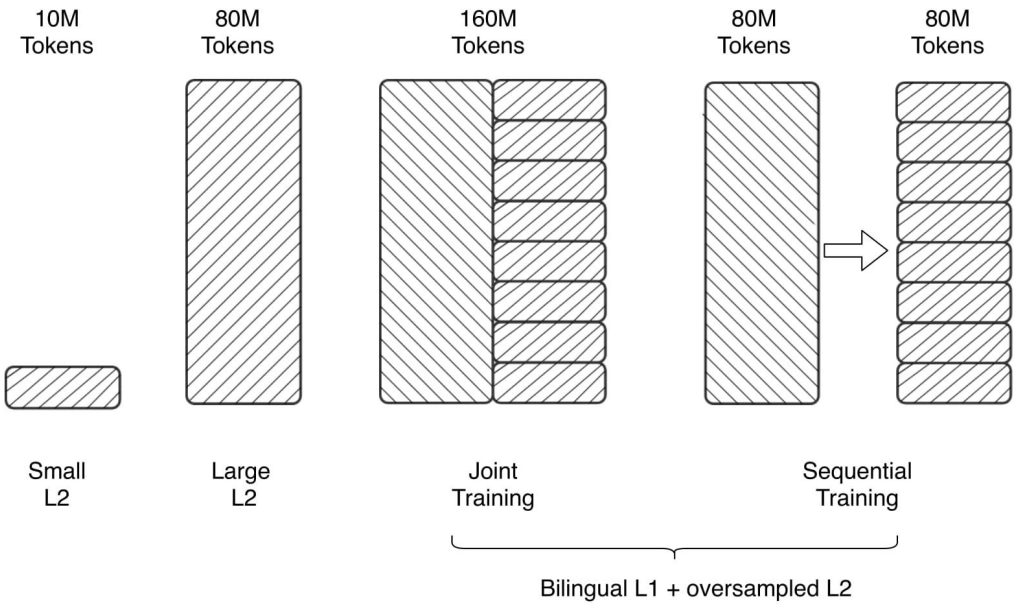
Natural Overlap

_ <FR> _Amsterdam _est _la _capitale .

_ <IT> _Amsterdam _è _una _città _oland ese .

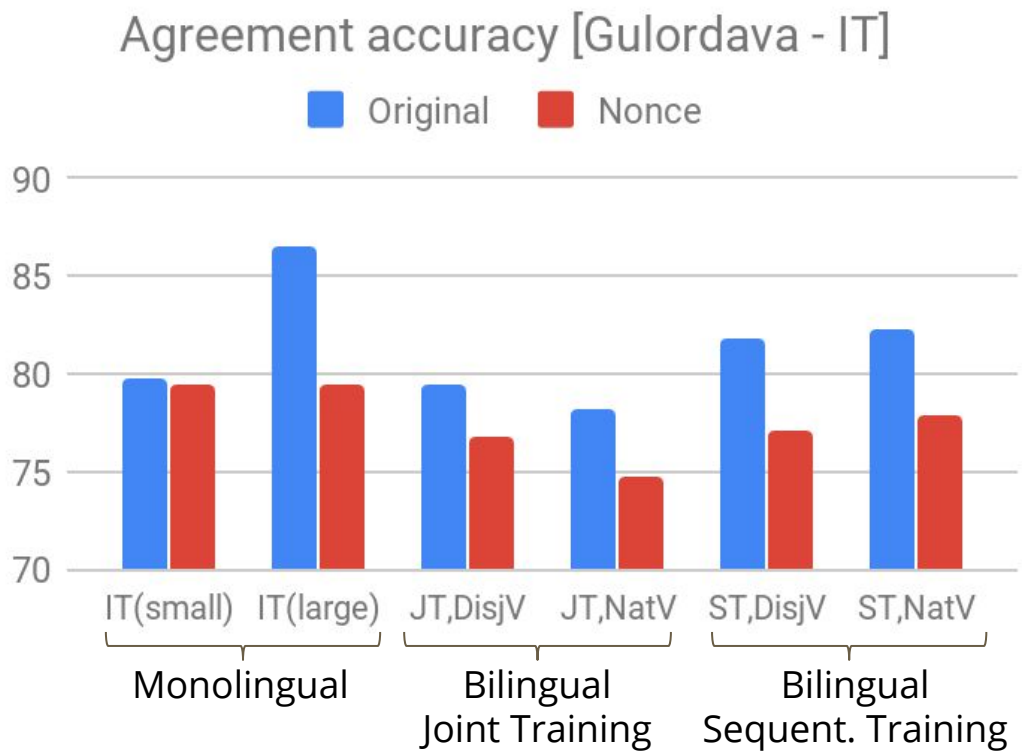
Sub-word Segmentation

Varying Conditions: 2) Training Regime



- Sequential training more efficient when adding new languages

Results



- Same trends in original vs nonce
- Training regime & Vocab overlap have both visible impact
- Train regime: Gains from helper language in sequential train
- Vocab overlap: Mixed effects
- No gains for nonce sentences

Limitations

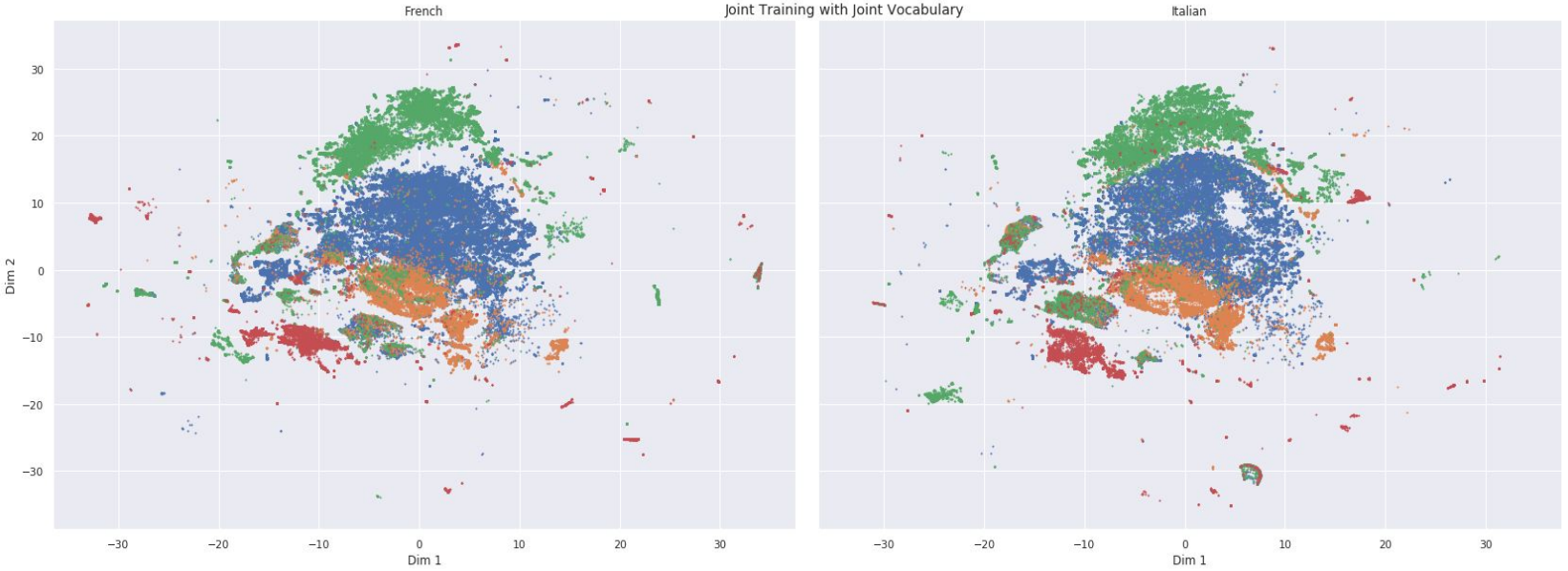
- Large performance variations across model initalizations
- Evaluation dataset is rather small
    - ~100 sentences for Italian (original)
- Task itself is complex
    - e.g. learn subject-verb agreement rules, learn lexical number categories, detect subject ...

- Taking a step back: Is there any sharing of syntactic categories?

Are syntactic categories shared?

Visualization

Joint Training and Natural Vocab Overlap



Hidden Layer 1

Nouns

Adjectives

Verbs

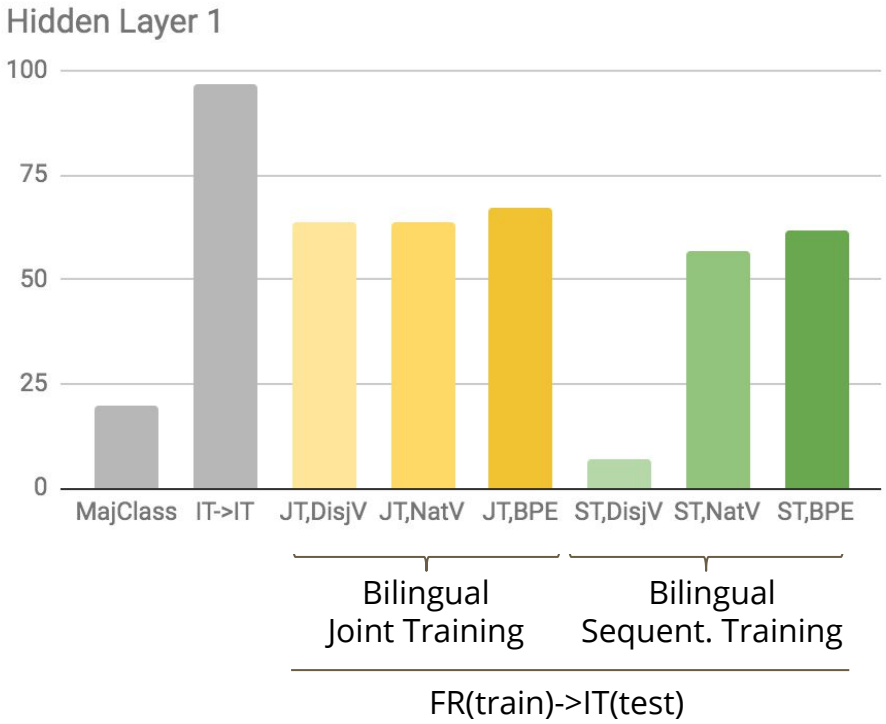
Adverbs

More Experiments

Quantify sharing of syntactic categories among languages:

- Diagnostic classifiers: POS, dependency relation type
- Train/test cross-lingually
- Linear classifiers
- Labels from Universal Dependencies

Results: PoS classification accuracy



- Transfer accuracies are low overall
- Joint training (slightly) better than sequential
- More vocab overlap = better transfer
- Overall small effects (except for outlier sequential train + disjoint vocab)
- Poor correlation with agreement accuracy results

Summary

- Cross-language transfer for complex syntactic tasks (long-range agreement) is limited, or none in purely grammatical conditions (nonce)
- At lower level, POS categories are shared to a limited extent
- [Next word prediction objective + Shared hidden layers] appear to be insufficient conditions for transfer
- Vocabulary overlap has slight and somewhat inconsistent impact

Future Work & Discussion

- Next steps: Look at different training objectives: translation, image-to-text
- Expectation: common output space (target language or visual) may give stronger incentive for transfer

- Which baseline for POS (or DepRel) classification..?
- What other tasks to probe syntactic transfer..?

THANKS FOR YOUR ATTENTION!

FEEDBACK?