The Role Of Common Ground For Referential Expressions In Social Dialogue

Jaap Kruijt & Piek Vossen

VU Amsterdam

INTRODUCTION

- In long-term human-agent relationships, efficient communication is key
- Over time communication evolves to be more efficient (e.g. Hawkins et al 2021)
- How can an agent learn and use this *common ground* to improve long-term interaction?

INTRODUCTION

- "Look at this old picture of *nana*…"
- Conventionalized references are hard to resolve!
  - Lots of knowledge assumed to be familiar
- What performance effect does this have for coreference resolution?

INNER CIRCLE VS OUTER CIRCLE

- *Inner circle*: part of common ground, conventionalized references
- *Outer circle*: individuals not mentioned before, only relevant for the current discourse
- How sensitive are machine learning models to this distinction?

	Inner	Outer
Nº individuals	50	**351**
Nº mentions	**2799**	2031

DATA

- Data-set of social dialogue with **temporal relations** between interactions based on episodes of *Friends* (Choi & Chen 2018)
- **Inner** / **outer** circle division
- Keeping the temporal structure intact, we developed **three** new train and development sets and **two** test sets
  - Test set A: *4 times as many* inner circle mentions *as* outer circle mentions (4/1 ratio)
  - Test set B: *equal amount of* inner circle and outer circle mentions (1/1 ratio)

DATA ANALYSIS

- Explorative analysis: do reference patterns for inner and outer circle differ?
- Inner circle: mostly **names** (NNP) and **sequences of names**
- Outer circle: mostly **pronouns** (PRP) and **sequences of pronouns**

	NNP	NN	PR	OTHER
Inner	**0.38**	0.24	0.30	0.08
Outer	0.26	0.24	**0.45**	0.05

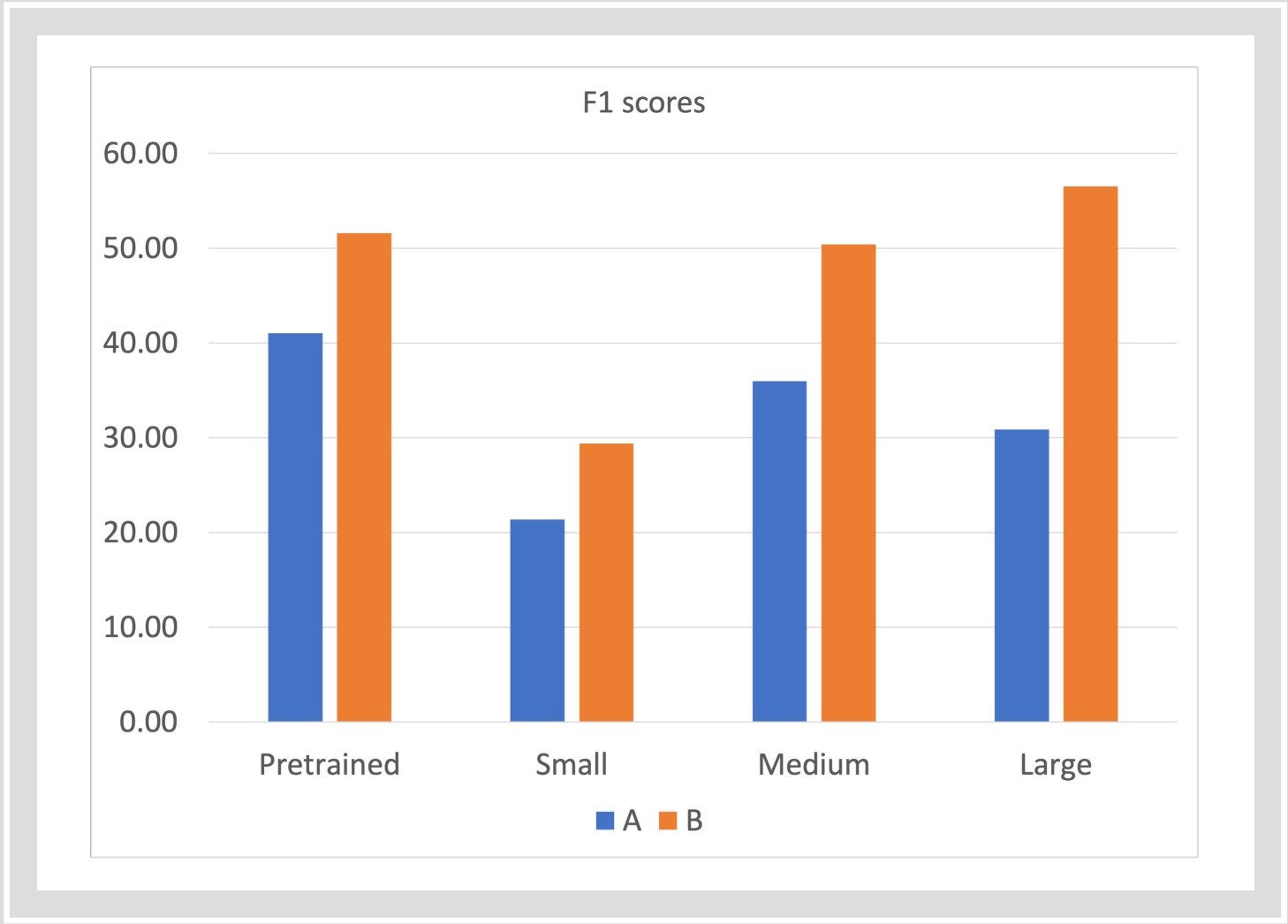
Inner circle	NNP	NN	PRP	*Outer circle*	NNP	NN	PRP
NULL	58.56	30.25	11.19	NULL	42.72	37.09	20.19
NNP	**52.98**	19.04	27.98	NNP	46.34	14.33	39.33
NN	24.35	44.50	31.15	NN	11.07	35.36	53.57
PRP	16.99	13.89	69.12	PRP	10.79	15.25	**73.96**

EXPERIMENT SETUP

- SpanBERT coreference resolution model (Joshi et al 2019)
- 4 models: pretrained, finetuned-small, finetuned-medium and finetuned-large
- Testing on both A (4/1 ratio) and B (1/1 ratio) to examine:
  - Performance on inner vs outer circle mentions
  - Effect of a larger prominence of inner circle mentions
  - Effect of finetuning on related background knowledge

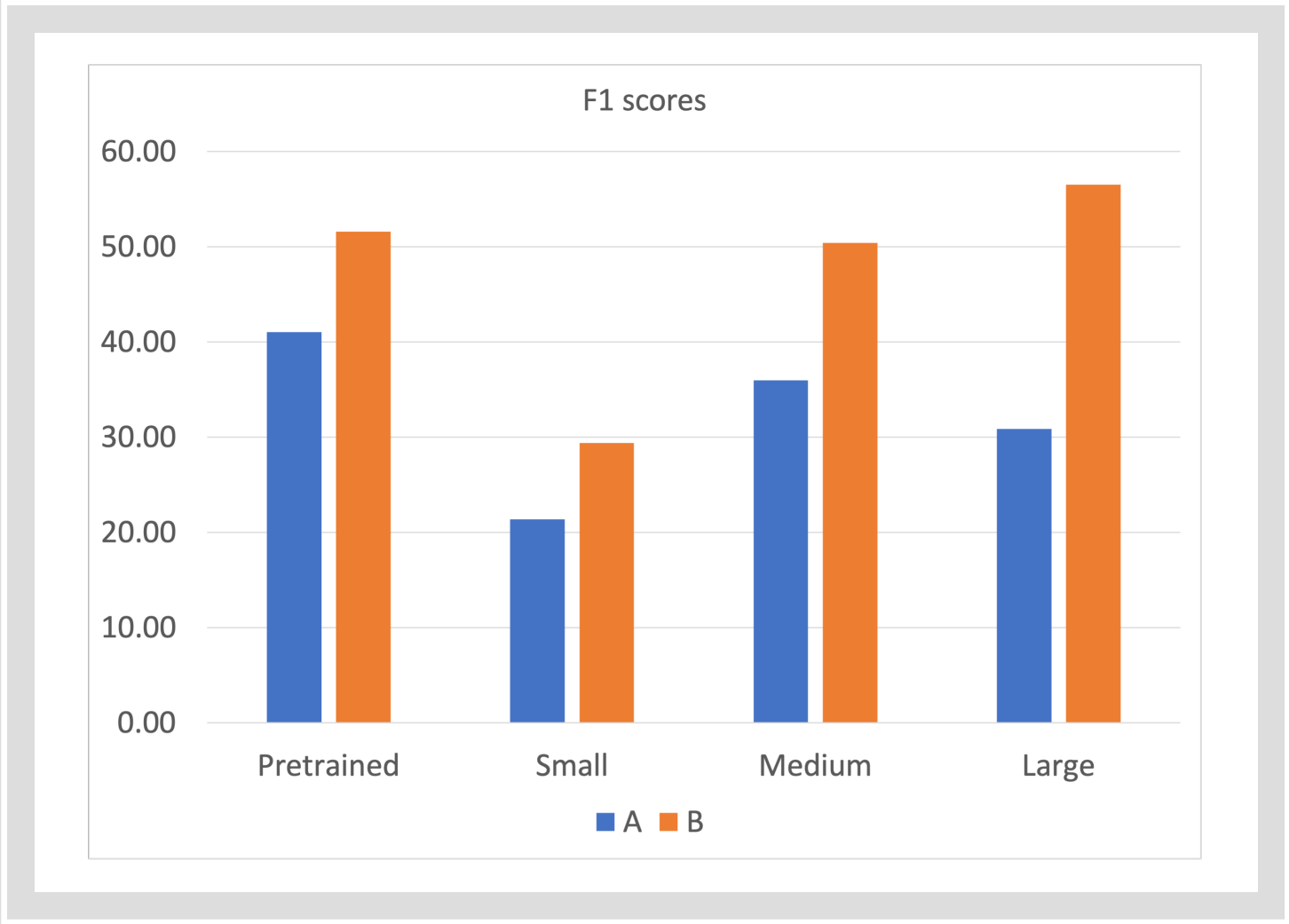
MODEL EVALUATION

- The **Pre-trained** model scores best for A (4/1 ratio)



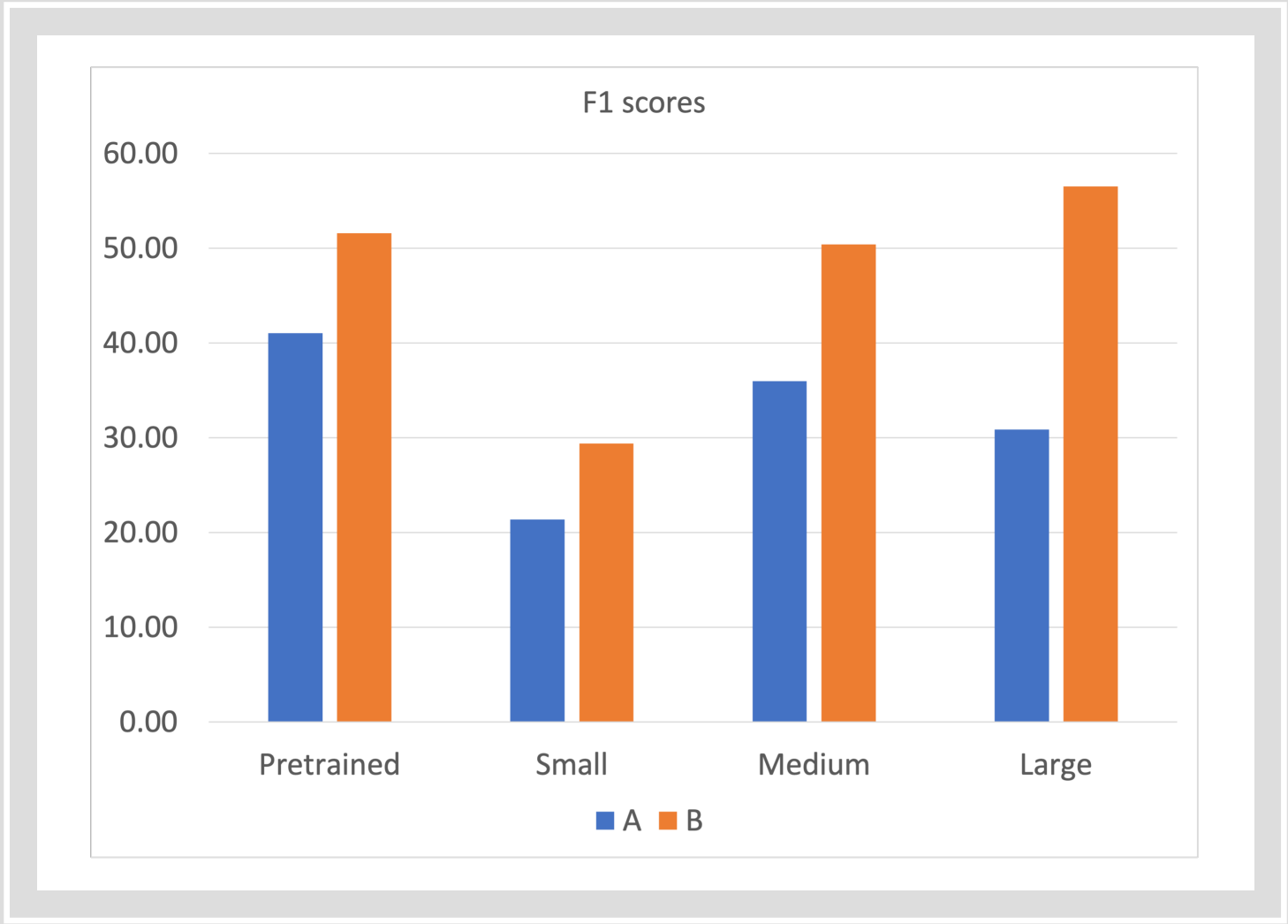
MODEL EVALUATION

- The **Pre-trained** model scores best for A (4/1 ratio)
- The **large** model scores best for B (1/1 ratio)



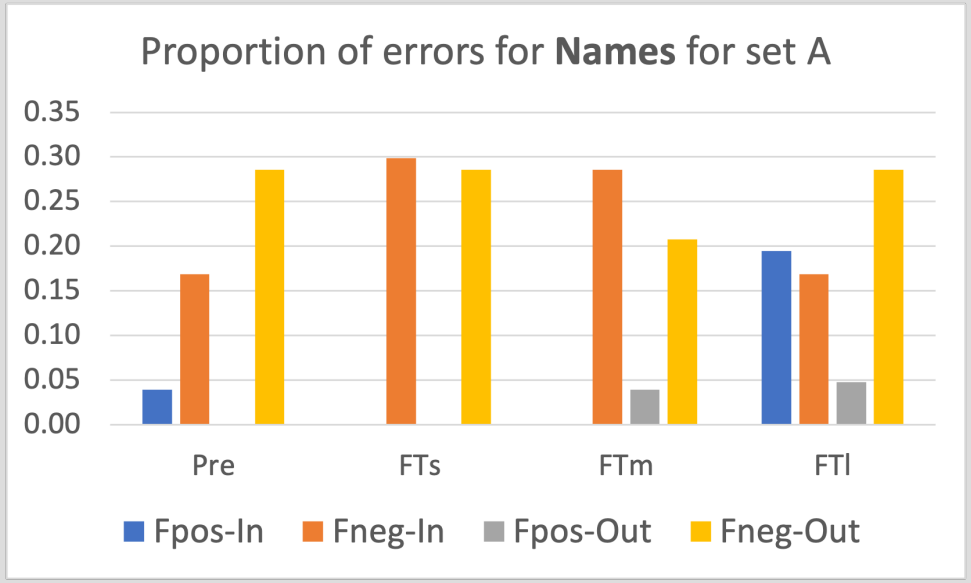
MODEL EVALUATION

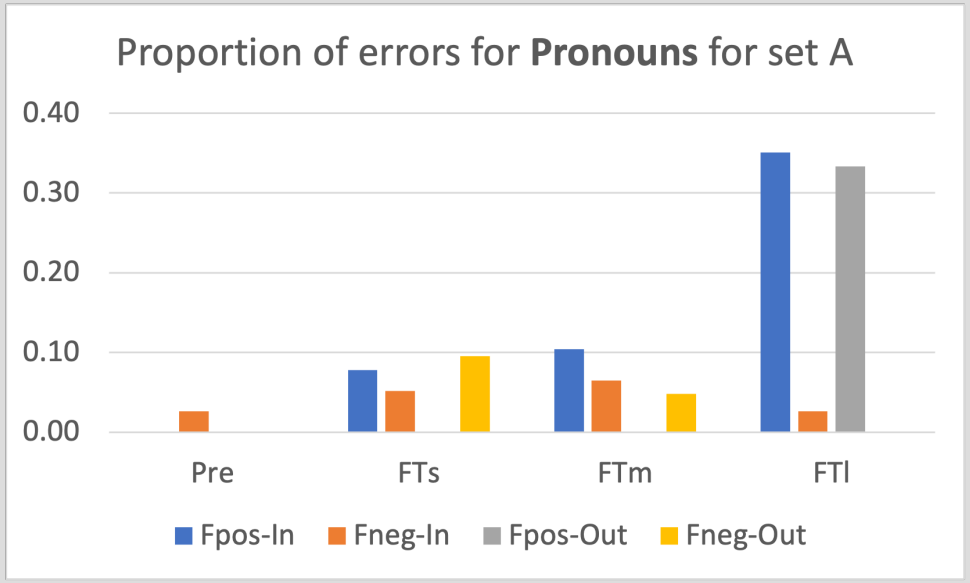
- The **Pre-trained** model scores best for A (4/1 ratio)
- The **large** model scores best for B (1/1 ratio)
- No real gain in fine-tuning for the inner circle mentions?



ERROR ANALYSIS

- More false negative errors for A vs more false positive errors for B
  - Strongest effect for the outer circle mentions
- Could point to trouble distinguishing between inner and outer circle
- More errors for **inner circle** mentions





ERROR ANALYSIS

- Most false positives with pronouns
- Most false negatives with names
- General **increase** of **false positives** associated with more fine-tuning
  - Notable increase for pronouns
- General **decrease** of false negatives

DISCUSSION

- Results and error analysis suggest that the model is probably over-fitting with more data
- Tendency towards learning more **discourse-related features** rather than learning from background knowledge
- No conclusions whether a knowledge-rich approach would be beneficial for this task

DISCUSSION

- We encourage more research into the role of common ground in reference resolution:
  - Data
  - Models
- We believe our approach and analysis are relevant for applications of agents which aim to establish a bond through shared social connections
- Future work will focus on a more knowledge-rich approach in an interactive setting

THANK YOU FOR YOUR ATTENTION!

Contact: j.m.kruijt@vu.nl

Code available at: https://github.com/cltl/inner-outer-coreference

REFERENCES

- Jinho D. Choi and Henry Y. Chen. 2018. SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64, New Orleans, Louisiana. Association for Computational Linguistics.
- Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2022). From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review.* Advance online publication.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Picture on slide 2 by Adam Winger on unsplash.com