We can use the **preference for reference pronoun translations** to make a **challenge dataset** and train an **evaluation measure**.

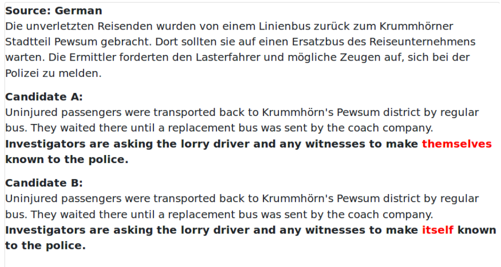
Evaluating Pronominal Anaphora: An Evaluation Measure and a Test Suite

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, Preslav Nakov

1 Motivation

- Monolingual/discourse-level evaluations of MT output reveal strong preference for human translations; BLEU scores are poor indicators of this
- Existing evaluations show low agreements with humans; targeted datasets are somewhat artificial

2 User Study & Dataset



Source: German
Die unverletzten Reisenden wurden von einem Linienbus zurück zum Krummhörner Stadtteil Pewsum gebracht. Dort sollten sie auf einen Ersatzbus des Reiseunternehmens warten. Die Ermittler forderten den Lasterfahrer und mögliche Zeugen auf, sich bei der Polizei zu melden.

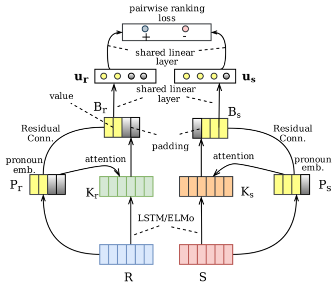
Candidate A:
Uninjured passengers were transported back to Krummhörn's Pewsum district by regular bus. They waited there until a replacement bus was sent by the coach company. **Investigators are asking the lorry driver and any witnesses to make themselves known to the police.**

Candidate B:
Uninjured passengers were transported back to Krummhörn's Pewsum district by regular bus. They waited there until a replacement bus was sent by the coach company. **Investigators are asking the lorry driver and any witnesses to make itself known to the police.**

- Create noisy data based on MT errors
- Conduct a user study to confirm that reference is better (>0.8 AC1 agreement)
- Source texts of errors form the **test suite for multiple source languages** (Source: WMT)

Source Language	Test Data from WMT Years	Unique Source Contexts
German	2011-2015,17	7,823
Czech	2011-2015,2017	6,713
French	2011-2015	4,659
Russian	2013,2014,2017	4,513
Spanish	2011-2013	4,417
Finnish	2015,2017	1,551
Turkish	2017	1,372
Hindi	2014	921
Chinese	2017	696
Latvian	2017	652

3 Model & Results



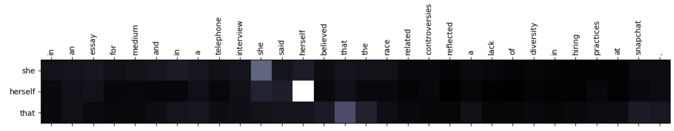
Exp	Context Setting	Test	Acc. (Glove)	Acc. (ELMo)
1	NC-Baseline	**R** vs. **R**′	69.12	85.80
2	NC	**R** vs. **R**′	68.97	88.04
3	NC	**R** vs. **S**	79.67	89.09
4	RC-Baseline	**R** vs. **R**′	69.07	85.80
5	RC	**R** vs. **R**′	67.88	87.90
6	CRC-Baseline	**R** vs. **R**′	69.16	86.66
7	CRC	**R** vs. **R**′	68.93	89.11
8	CRC	**R** vs. **S**	77.87	90.69

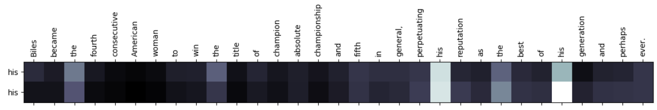
Language	Acc.(ELMo)	AC1 Agr.
Russian→English	79.4	0.80
French→English	82.0	0.84
German→English	81.6	0.83
Chinese→English	82.4	0.83
- - Only English	—	0.83
Overall (average)	81.35	—

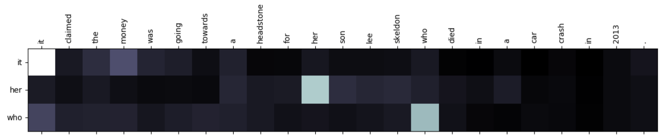
- Distinguish good from bad pronoun translations: pairwise ranking loss training with reference vs. MT
- Helps to include common reference context; results on noisy data indicative of the model's sensitivity to pronouns
- Model predictions agree (>0.8) with humans
- Attention heat maps show the model identifies wrong pronouns despite no specific signal; gives greater score to animacy/consistency

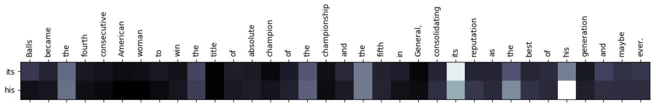
4 Future Work

- Handle multiple suitable pronouns and other discourse phenomena









SCAN ME

← Download paper/code/data
https://ntunlpsg.github.io/project/discomt/eval-anaphora/

NTU NLP