Master's thesis defense

Neural models for multilingual natural language understanding and generation

Nalin Kumar

Thesis supervisor: Mgr. et Mgr. Ondřej Dušek, Ph.D.

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Motivation

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees

(Trane | foundingDate | 1913-01-01),
(Trane | location | Ireland),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | numberOfEmployees | 29000)

- Main focus on text-to-data and data-to-text generation

Motivation

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees

(Trane | foundingDate | 1913-01-01),
(Trane | location | Ireland),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | numberOfEmployees | 29000)

- Main focus on text-to-data and data-to-text generation
  - Effective way of structuring natural text and generating paraphrases

Motivation

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees

(Trane | foundingDate | 1913-01-01),
(Trane | location | Ireland),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | numberOfEmployees | 29000)

Trane is based in Ireland and was founded on January 1st 1913 in La Crosse, Wisconsin. It has 29,000 employees

(Trane | location | Ireland),
(Trane | foundingDate | 1913-01-01),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | numberOfEmployees | 29000)

Trane, having 29,000 employees, was founded on January 1st 1913 in La Crosse, Wisconsin. It is based in Ireland.

(Trane | numberOfEmployees | 29000),
(Trane | foundingDate | 1913-01-01),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | location | Ireland)

Motivation

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees

(Trane | foundingDate | 1913-01-01),
(Trane | location | Ireland),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | numberOfEmployees | 29000)

- Main focus on text-to-data and data-to-text generation
  - Effective way of structuring natural text and generating paraphrases
  - Can be further used for extracting/removing (ir)relevant information

Motivation

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees

(Trane | foundingDate | 1913-01-01),
(Trane | location | Ireland),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | numberOfEmployees | 29000)

- Main focus on text-to-data and data-to-text generation
  - Effective way of structuring natural text
  - Can be further used for extracting/removing information
  - Can also be used for checking similarity between two texts (or even generating fact-checked outputs)

Motivation

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees

(Trane | foundingDate | 1913-01-01),
(Trane | location | Ireland),
(Trane | foundationPlace | La_Crosse,_Wisconsin),
(Trane | numberOfEmployees | 29000)

Trane was founded on January 1st 1915.

(Trane | foundingDate | 1915-01-01),

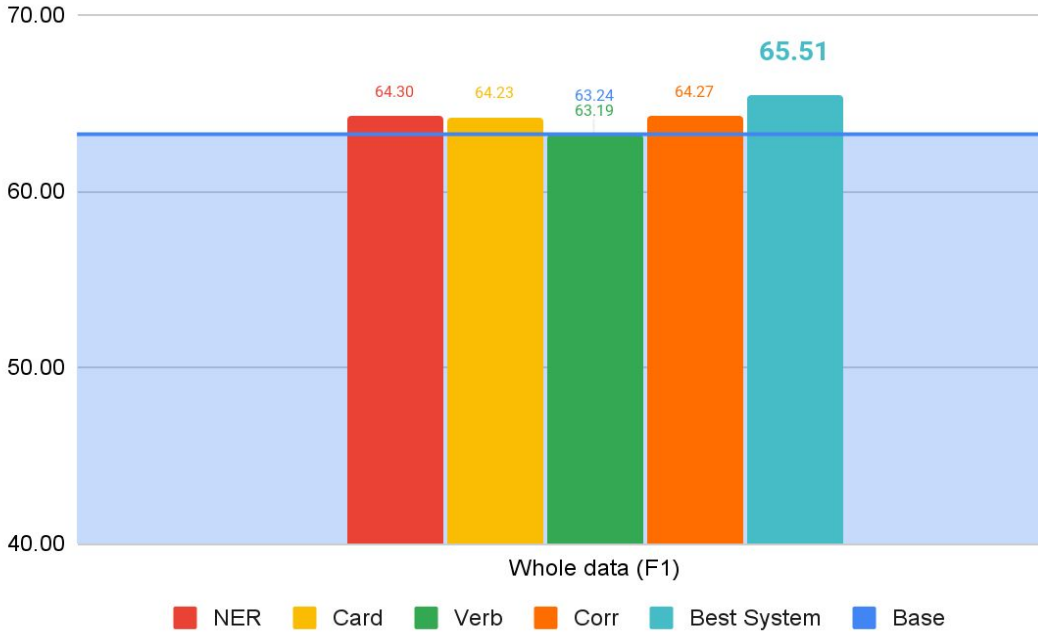
Introduction

- RDF — Resource Description Framework — represents data on web
- Information stored in graphs containing set of triples
  - Triple — two nodes representing subjects and objects, and relation between them
- Tasks: Natural text ↔ Set of RDF Triples
- Dataset: WebNLG 2020 & 2023
  - 2020: English & Russian (~35k & ~14k Lexicalisations)
  - 2023: Introduced 4 low-resourced languages — Breton, Irish, Maltese & Welsh
    - Training set created using Edinburgh Zero system (Zhang et al. [2020])
    - Poor quality — Incomplete translations

RDF parsing

- Task: Text-to-RDF triples generation for English
- Various auxiliary tasks — NER, cardinality prediction, RDF-to-text generation, etc.
- Most useful — Cleaning training data — Aligning the triples with verbalisations (align-then-generate)
    - Order of RDF triples might be different to the entities in verbalisations
    - Attention scores to find correspondence
- Baseline: T5-small finetuned on the WebNLG data (2020)
- Evaluation Metric: Strict Match (F1)

Results



- All methods perform better than baseline — improvement not so substantial
- Baseline has decent outputs — struggles on unseen
- *align-then-generate* — better generalization among all
  - Covers more triples
- Other systems — similar to baseline — either miss out triples or struggle with named entities

Summary

- Joint training for multiple tasks can generalize better on unseen categories.
- In general, performance difference between baseline and other models is not big — auxiliary tasks are probably not “intelligent” enough or model is saturated for the given dataset

Multilingual Semantic Parsing

- Semantic parsing in multilingual setting — Multilingual text to English RDF triples
- Model: mT5 base
- Dataset: WebNLG 2023* (English (En), Irish (Ga), Maltese (Mt), Welsh (Cy), Russian (Ru))
- Model Variants:
  - **Baseline**
    - *Resynthesized training data (WebNLG 2020) using better translation system (NLLB)
    - Irish (Ga), Maltese (Mt), Welsh (Cy)
    - Finetuned mT5 on individual languages
  - **Multilingual model (MLM)**

Results

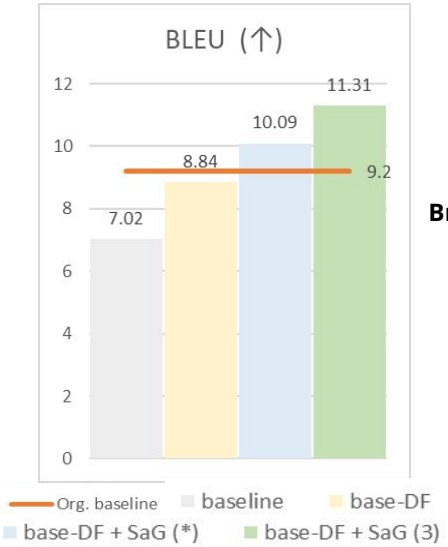
Lang	Model	Whole data (F1)
En	Align	**65.51**
	Base	62.94
	MLM	50.40
Ga	Base	43.64
	MLM	**44.63**
Mt	Base	**54.51**
	MLM	48.39
Cy	Base	**49.57**
	MLM	46.23
Ru	Base	**85.85**
	MLM	75.09

- For En — earlier method works better — T5 is English focused
- Base models perform better — other than Ga
- MLM for Ga — least resourced in mT5 training data — cross-linguality helps
- The models struggle with numbers and longer verbalisations

Multilingual RDF-to-Text Generation

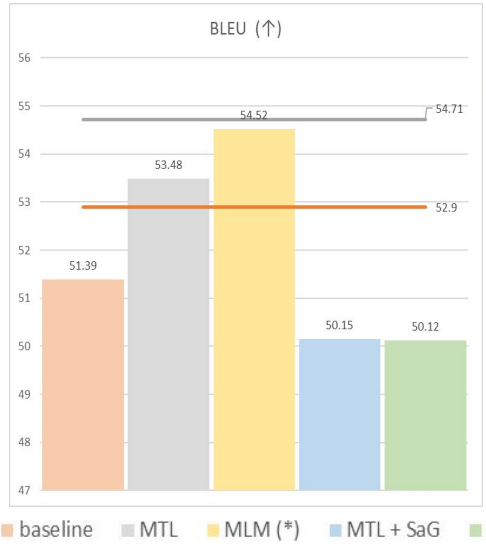
- Task: RDF-triples to text
- Languages: Breton (Br), Irish (Ga), Maltese (Mt), Welsh (Cy), Russian (Ru)
- Challenges (except Ru):
  - Small and inconsistent training data
  - Limited additional resources
- Proposed approaches:
  - Employ multilingual T5 model (**base**)
  - Modify training data: better NMT system & filtering
    - Ga, Mt, Cy — Similar to previous (Using NLLB) (**base-NLLB**)
    - Br — NLLB has no support for Br — Filtered out bad samples
  - Using additional languages/tasks (**MLM**/**MTL**)
  - Split and generate (**SaG**) — split triples and generate — more verbose

Results — Br & Ru



Breton

- Data filtering works
- SaG brings further improvements in scores

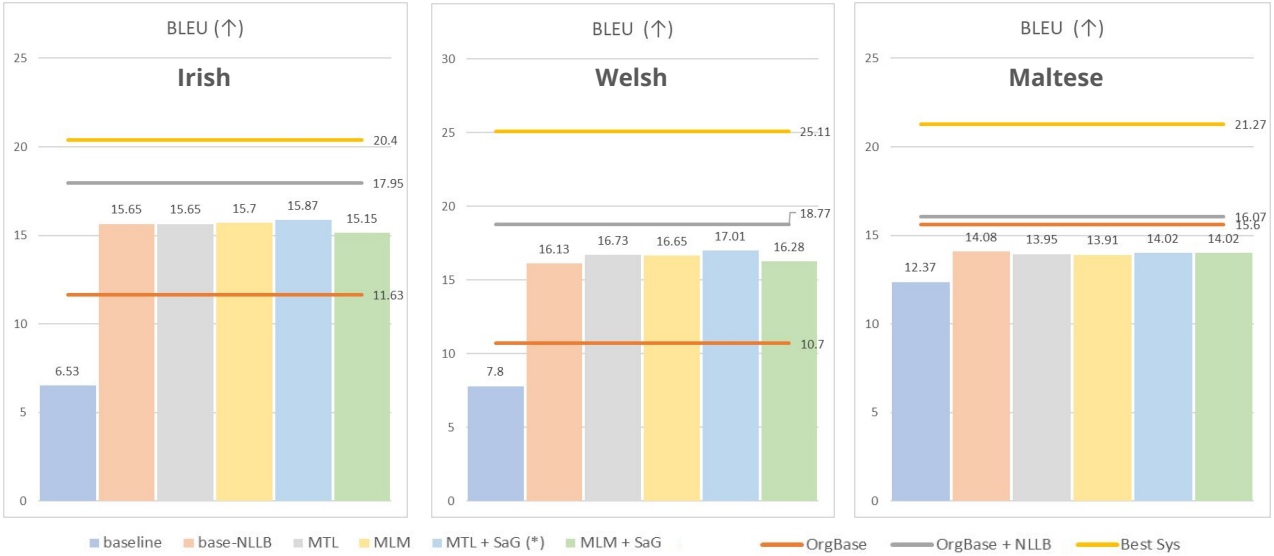


Russian

- Multilingual model — improved performance
- SaG does not work

(*) Submitted system

Results — Ga, Cy, Mt



(*) Submitted system

- Significant jump from baseline to base-NLLB
- No further huge improvements

Summary

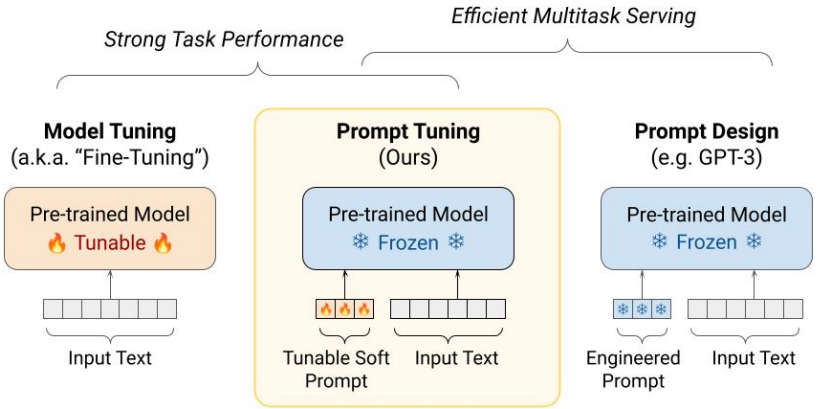
- No universal solution — different solutions work for different languages
- Better NMT system in generate-and-translate — boosts performance
- Modifying training data improves the results significantly
- SaG — low-resource languages 👍 — Ru 👎

Is tuning all model parameters really necessary?

- Effectiveness of finetuning all parameters VS prompt tuning small # params
  - Task: RDF Triple to Text Generation for En only
  - WebNLG 2020 dataset
- Prompt tuning — efficient model tuning
- Variables — Model sizes, types and training sample size
- Models: T5 & Flan-T5

Is tuning all model parameters really necessary?

- Effectiveness of finetuning all parameters VS prompt tuning small # params
- Prompt tuning — efficient model tuning
  - Initialize fixed length vectors (soft prompt) (our case: 100)
  - Prepend to embedded input
  - Tune these parameters
- Variables
- Models: T5 & Flan-T5



Img src: https://research.google/blog/guiding-frozen-language-models-with-learned-soft-prompts/

Is tuning all model parameters really necessary?

- Effectiveness of finetuning all parameters VS prompt tuning small # params
- Prompt tuning — efficient model tuning
- Variables — Model sizes, types and training sample size
  - Model Size — small, base, large
  - Training data size — 0.5k, 2.5k, 10k, whole data
- Models: T5 & Flan-T5

Results

		small		base		large	
		FT	PT	FT	PT	FT	PT
T5	0.5k	31.73	31.72	39.3	39.78	44.44	45.6
	2.5k	35.15	34.21	46.31	44.84	47.97	47.66
	10k	47.54	33.39	50.55	46.25	52.7	48.81
	whole	49.78	34.34	52.23	45.06	53.41	49.29
Flan-T5	0.5k	39.34	40.69	42.27	49.59	46.8	52.49
	2.5k	40.47	41.98	48.5	49.93	47.78	52.51
	10k	40.64	41.79	45.66	49.97	47.59	52.98
	whole	40.16	42.36	45.54	50.24	47.06	53.54
ZS		48.27		50.49		54.29	

BLEU Scores
FT: Finetuned, PT: Prompt-tuned, ZS: Flan-T5 Zero Shot

- T5
  - Comparable performance for lower sample size
  - Training sample ↑ BLEU ↑
    - PT — increase less substantial — underfitting?
  - Model size ↑ difference in performance of FT and PT ↓
- Flan-T5
  - Training sample ↑ No specific pattern
    - Performance decline in several cases (FT – more and PT – less)
    - Training data already includes WebNLG
    - Thus potential overfitting

Is tuning all model parameters really necessary?

Conclusion

- Major focus on Text-to-RDF parsing
  - Joint approaches work better, but does not give a huge boost
- Multilingual Text-to-RDF parsing
  - Finetuning multilingual model on individual languages – better
- Multilingual RDF-to-Text generation
  - Splitting triples and generating verbose verbalisations work for lower-resourced languages
- Parameter-efficient tuning
  - Prompt tuning can give comparable performance to finetuning for larger models or smaller datasets

Thank you

Contact Me @

Nalin Kumar
knalin55@gmail.com
https://knalin55.github.io

knalin55

Inclusion of spurious entities in WebNLG evaluation.

Reference: <John, play, football>

Candidate: <John, play, hockey>

(football, obj) — (hockey, obj): Spurious

- Essential for calculating precision
- Also, a measure of hallucination

Back

Combining promising approaches for RDF parsing.

- No substantial improvement — training saturation
  - align-then-generate — data modification method — combination easy with other methods
  - Combined with corr — Strict F1: 64.97 Best: 65.51

Back

Discrepancy between the METEOR scores and the other metrics in Table 5.3

- Error from my side.
- METEOR does not support for given languages.

Model	BLEU	METEOR	ChrF	TER
Lorandi and Belz [2023]	25.11	-	0.55	0.64
Guo et al. [2020b] + Edin	10.70	-	0.36	0.77
Guo et al. [2020b] + NLLB	**18.77**	0.25	**0.48**	**0.70**
baseline	7.80	0.15	0.29	0.78
base-NLLB	16.13	0.24	0.44	0.79
MTL	16.73	0.24	0.45	0.78
MLM	16.65	0.24	0.44	0.80
MTL + SaG ◇	17.01	0.24	0.45	0.79
MLM + SaG	16.28	**0.35**	0.45	0.81

Welsh (*Cy*) – 3rd of 4 submitted systems in the WebNLG 2023 Challenge

- Other scores are calculated using METEOR-other, while the highlighted one is done using the default METEOR (used for En) (Will fix it)

Back

Joint training of the text generation with the RDF triple parsing in Chapter 5 (similarly to what you did in Chapter 4).

- MTL — SP + TG + translation task — Ru and Mt ↓, Cy ↑ & Ga (no difference)
- For En (TG + SP), the BLEU score slightly decreases from 31.25 (TG) to 30.70 (TG + SP)

Back