WMT2023, Singapore, 7.12.2023

Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies

Kirill Semenov[C] Vilém Zouhar[E] Tom Kocmi[M] Dongdong Zhang[M]
Wangchunshu Zhou[A] Yuchen Eleanor Jiang[A]

[C]Charles University [E]ETH Zürich [M]Microsoft [A]AIWaves

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

Motivation and Task

- For specific domains, *accurate* and *consistent* terminologies are critical
- No widely used metrics or solutions
- Terminologies are easier to collect

Assess the extent to which an MT system can:

- Make use of the additionally provided dictionary
- Adhere to the specific terminology constraints

Three Modes

Base	**Source**: “Most informative is the analysis of airway secretions:” **Reference**: “Häufig jedoch führt die Analyse von Material aus den Atemwegen zur Diagnose:” **Terms**: {}
Proper	**Source**: “Most informative is the analysis of airway secretions:” **Reference**: “Häufig jedoch führt die Analyse von Material aus den Atemwegen zur Diagnose:” **Terms**: {“analysis of airway secretions” → “Analyse von Material aus den Atemwegen”}
Random	**Source**: “Most informative is the analysis of airway secretions:” **Reference**: “Häufig jedoch führt die Analyse von Material aus den Atemwegen zur Diagnose:” **Terms**: {“Most”→“Häufig”}

Setup

Data: Languages, Domains and Annotation

Chinese→English:

- BWB corpus (Jiang et al. 2023); web novels domain
- Manual terminology annotation

X→Y	Count	X/Y Words	Terms
German→English	2963	22.2/22.6	3.8
English→Czech	3005	25.6/21.6	3.6
Chinese→English	2640	9.7/36.9	1.1

German→English:

- MuchMore Springer Bilingual Corpus; medical papers
- GPT-4 terms extraction + human post-editing

English→Czech:

- Czech and English abstracts of ÚFAL papers (Rosa&Zouhar, 2022); NLP abstracts
- GPT-4 terms extraction + human post-editing

Metrics

General accuracy: chrF, COMET*, COMET-KIWI*

Consistency: by Semenov&Bojar, 2022.

- Reference-less metric
- Compares each term's translations to its first translation
- Lemmatized

Success rate:

- Regex / fuzzy match, *surface tokens* / lemmas

Comparison with Previous Run

Difference in setups between Terminology **WMT2021** and **WMT2023**:

1. language pairs and domains:
   - 1.1. En→{Fr, Zh, Ru, Ko}, Cs→De VS {Zh, De}→En, En→Cs
   - 1.2. Medical (COVID-19) VS medical (general), web novels, academic

2. Annotation:
   - 2.1. Term extraction: human VS GPT4+human, human
   - 2.2. Modes: terms VS proper terms, random terms, no terms

3. Terminology metrics:
   - 3.1. Reference-based success rate+consistency VS reference-based success rate, reference-less consistency

Participants

Participants: Overview

- 7 participants, 15 submitted systems
- Language pairs coverage:
  - zh-en: 15/15 systems
  - en-cs, de-en: 7/15 systems
- Main approaches:
  - Source-based:
    - Terminology injection
    - Copy mechanism, separate encoders (src, terminology)
  - Target-based:
    - Constrained decoding
    - Post-editing (incl. LLMs)
  - Synthetic data: sentences with terminology; unsupervised terminologies

Results and Discussion

Results: Overview - NEW!

System	ChrF De→En	ChrF En→Cs	ChrF Zh→En
AdaptTerm	61.0	64.4	37.5
Lingua Custodia	61.8	67.7	32.6
OPUS-CAT	68.3★	75.1★	27.7
UEDIN$_{LLM}$	60.0	64.8	41.2
UEDIN$_{Tag}$	58.3	64.7	41.0
UEDIN$_{Twoshot}$	60.5	62.4	34.5
BJTU-LB			43.8★
VARCO-MT$_{TSSNMT}$			43.0
VARCO-MT$_{ForceGen}$			40.5
Huawei	62.1	58.2	36.8

System	COMET$^{DA}_{22}$ De→En	COMET$^{DA}_{22}$ En→Cs	COMET$^{DA}_{22}$ Zh→En
AdaptTerm	0.801	0.841	0.688
Lingua Custodia	0.735	0.834	0.609
OPUS-CAT	0.828★	0.889★	0.557
UEDIN$_{LLM}$	0.813	0.869	0.757★
UEDIN$_{Tag}$	0.809	0.868	0.757★
UEDIN$_{Twoshot}$	0.792	0.835	0.650
BJTU-LB			0.751
VARCO-MT$_{TSSNMT}$			0.755
VARCO-MT$_{ForceGen}$			0.715
Huawei	0.843	0.887	0.666

System	Terminology Consistency De→En	Terminology Consistency En→Cs	Terminology Consistency Zh→En
AdaptTerm	0.617	0.753	0.750
Lingua Custodia	0.602	0.766	0.696
OPUS-CAT	0.661★	0.808★	0.293
UEDIN$_{LLM}$	0.588	0.741	0.713
UEDIN$_{Tag}$	0.606	0.750	0.755
UEDIN$_{Twoshot}$	0.574	0.737	0.622
BJTU-LB			0.764
VARCO-MT$_{TSSNMT}$			0.971
VARCO-MT$_{ForceGen}$			0.773★
Huawei	0.788	0.603	0.562

System	Terminology Success Rate De→En	Terminology Success Rate En→Cs	Terminology Success Rate Zh→En
AdaptTerm	0.591	0.577	0.785
Lingua Custodia	0.632	0.640	0.774
OPUS-CAT	0.948★	0.932★	0.133
UEDIN$_{LLM}$	0.557	0.594	0.750
UEDIN$_{Tag}$	0.532	0.584	0.765
UEDIN$_{Twoshot}$	0.560	0.498	0.452
VARCO-MT$_{TSSNMT}$			0.779
VARCO-MT$_{ForceGen}$			0.793★
BJTU-LB			0.759
Huawei	0.690	0.455	0.529

Best Performers

De-En, En-Cs: OPUS-CAT, Lingua Custodia, AdaptTerm, UEDIN-LLM

Zh-En: BJTU-LB, Varco MT (ForceGen), UEDIN-LLM, UEDIN-Tag

- All approaches work
- Zh VS others
- chrF, COMET - quality improves with *any* dictionary
  - consistency and success rate react more to proper terminology
- System ranking similar for chrF, COMET and term success rate, but differ for term consistency

Results: Average Difference - NEW!

System	ChrF		$COMET_{22}^{DA}$		T. Consistency		T. Success Rate	
	+Proper	+Random	+Proper	+Random	+Proper	+Random	+Proper	+Random
AdaptTerm	9.0	11.6	0.043	0.054	0.020	-0.010	0.3	0.338
Lingua Custodia	10.1	11.8	0.032	0.026	0.118	-0.016	0.402	0.369
OPUSCAT	10.2	9.2	0.031	0.043	0.055	0.187	0.345	0.247
$UEDIN_{LLM}$	6.4	7.5	0.011	0.017	0.027	0.018	0.214	0.157
$UEDIN_{Tag}$	5.4	6.5	0.010	0.013	0.055	0.009	0.218	0.127
$UEDIN_{Twoshot}$	6.9	5.9	0.029	0.012	0.045	-0.013	0.193	0.165
BJTU-LB †	2.5	0.8	0.015	0.007	0.058	0.049	0.252	-0.208
$VARCO\text{-}MT_{TSSNMT}$†	8.3	4.7	0.054	0.017	0.171	0.089	0.515	-0.041
$VARCO\text{-}MT_{ForceGen}$ †	3.4	0.9	0.019	0.003	0.166	0.021	0.529	-0.106
Huawei	0.2	0.9	-0.004	0.010	-0.010	-0.090	0.038	-0.113

- ChrF, COMET - any terminology helps
- Consistency, success rate - improvement on proper terminology

Discussion, Limitations and Perspectives

Discussion:

- Low correlation of consistency VS any other metric:
  - Pay more attention for competitors?
  - Or improve a metric?

Limitations:

- Not enough controlled parameters (incl style/domain, terminology extraction and systems applied to all languages)
- No qualitative analysis

Perspectives:

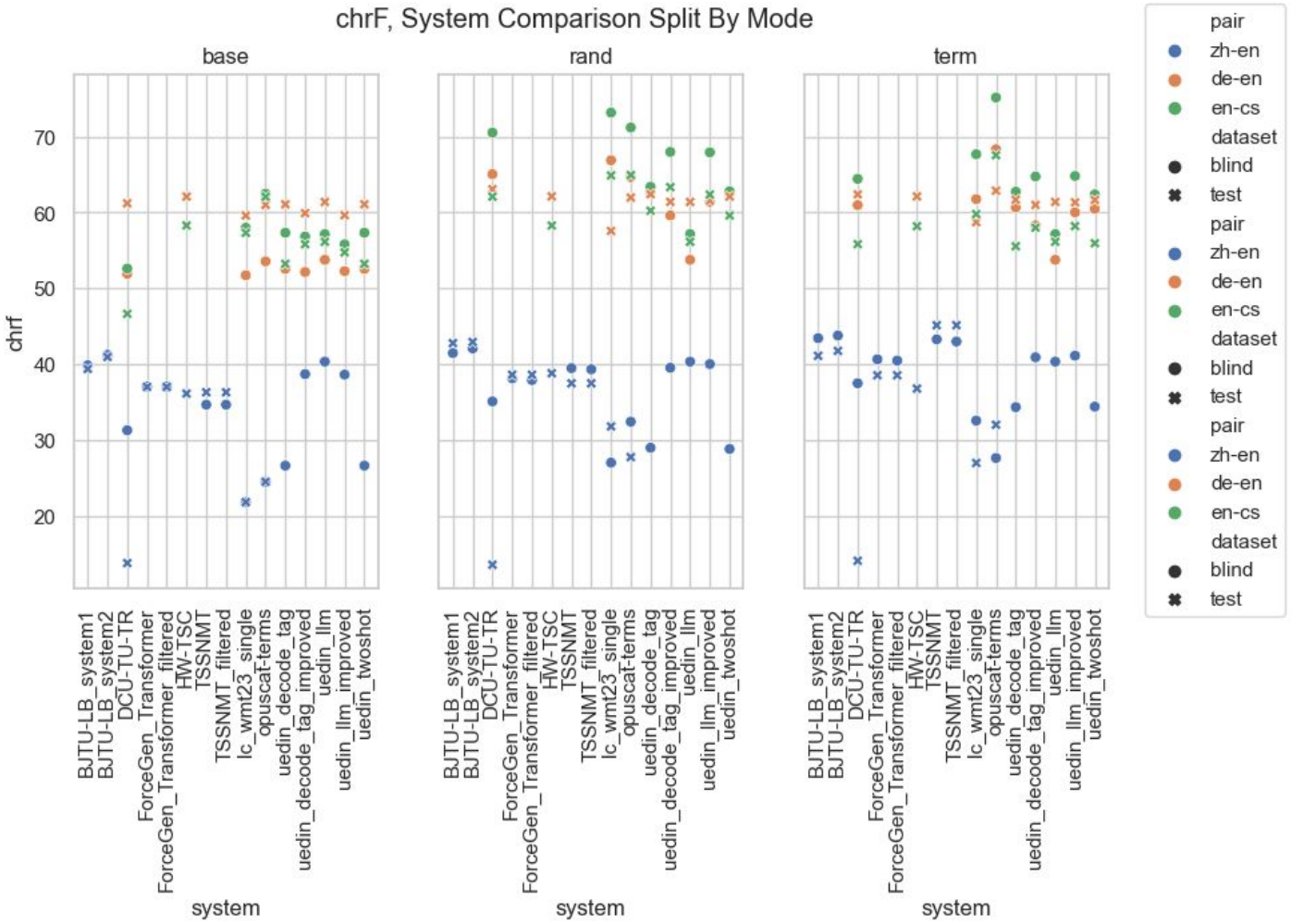
- Replication and more setup consistency over years?
- Other languages - why Zh-En is so different?

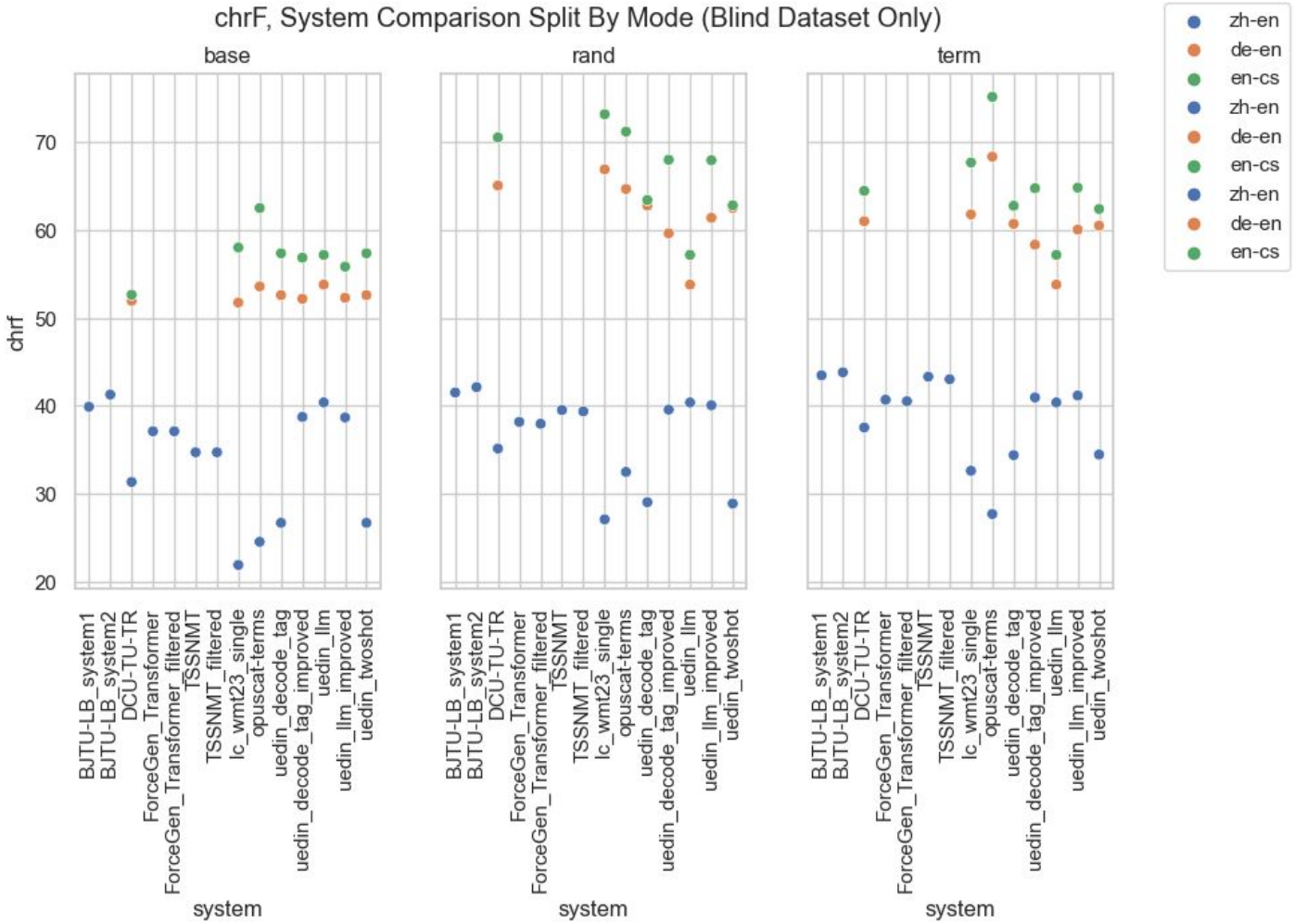
Thank you!

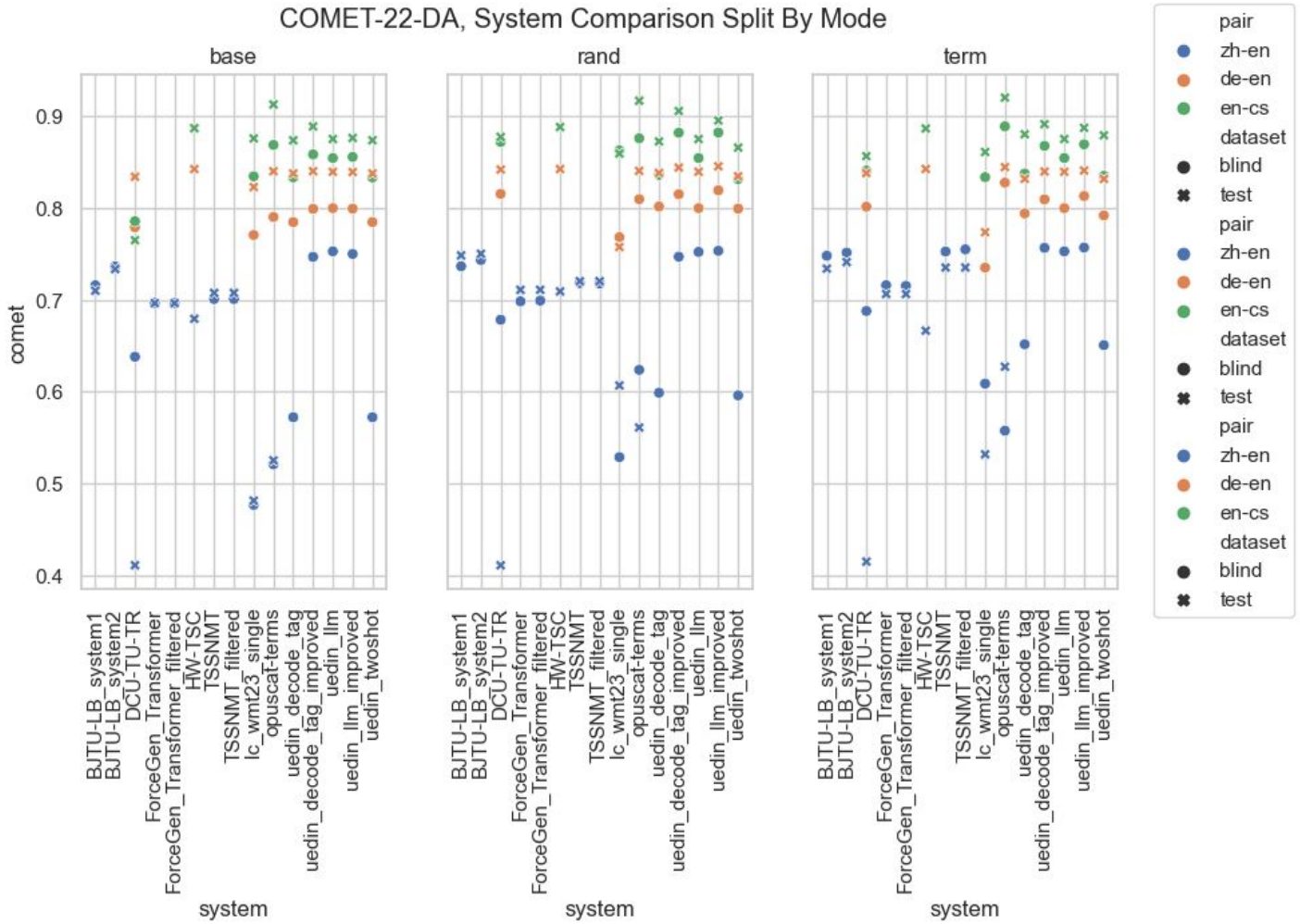
All updated statistics are available at the Shared Task web page:

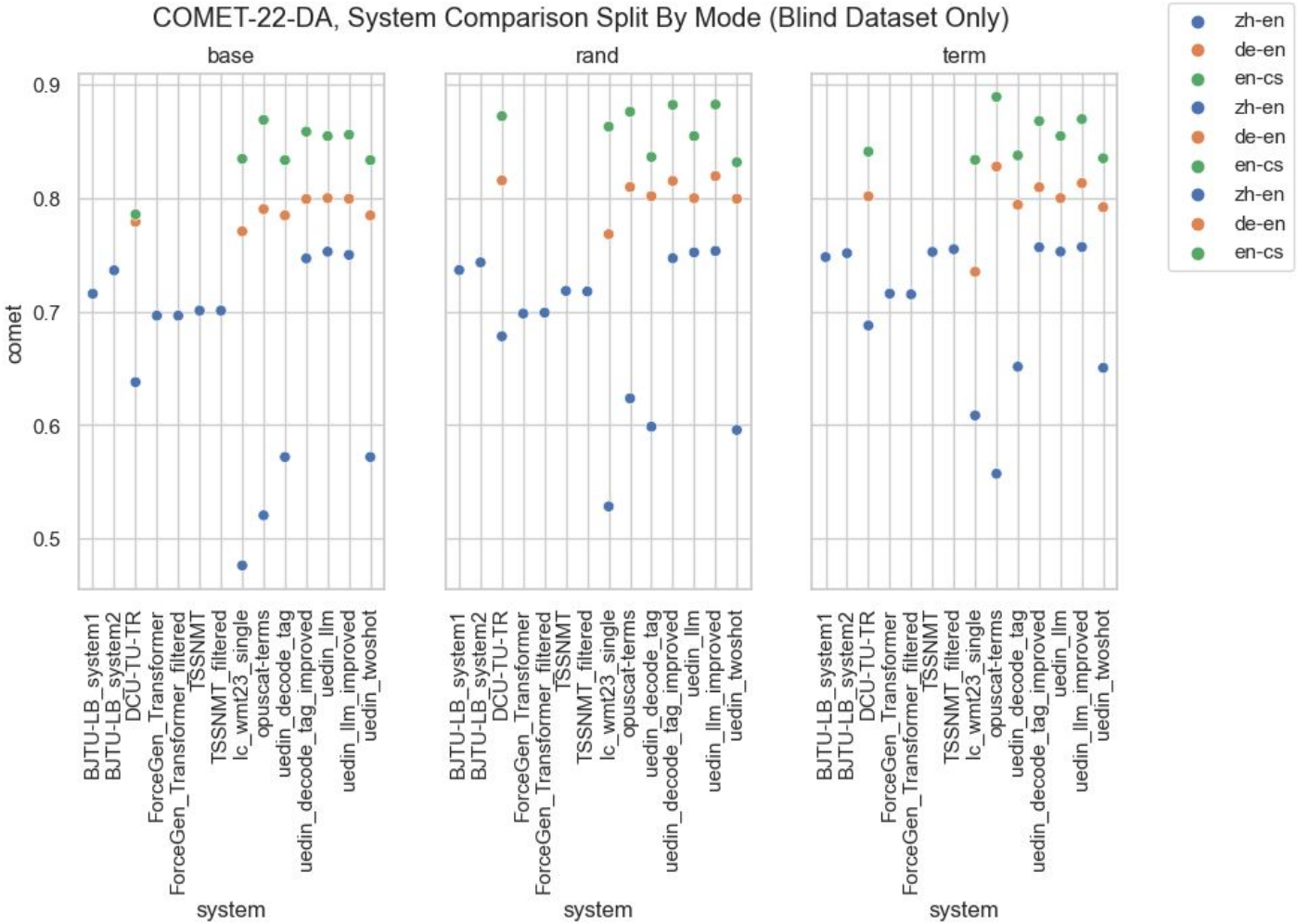
wmt-terminology-task.github.io/

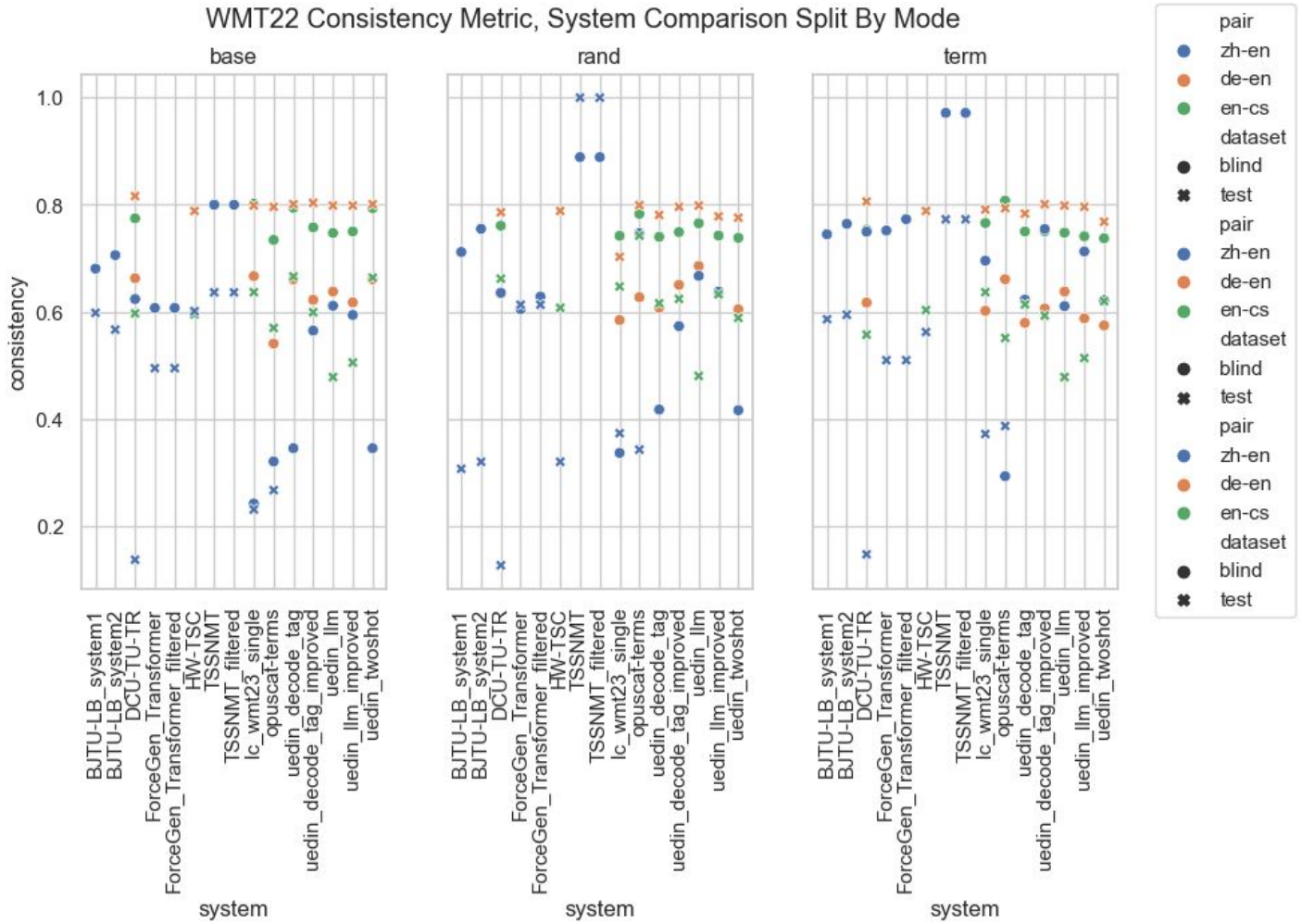
Additional Slides

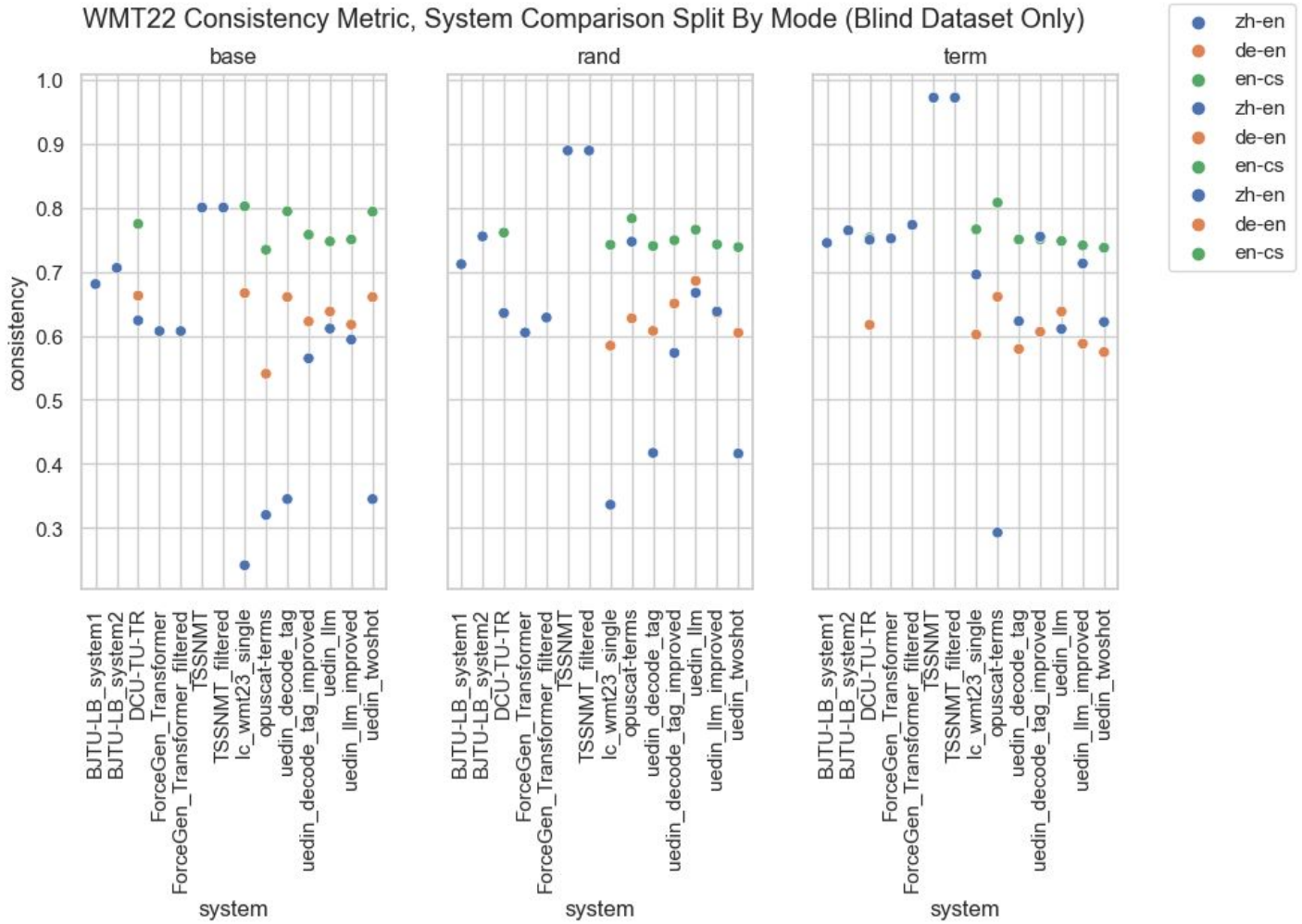

chrF, System Comparison Split By Mode
base
rand
term
chrf
70
60
50
40
30
20
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
HW-TSC
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
pair
zh-en
de-en
en-cs
dataset
blind
test

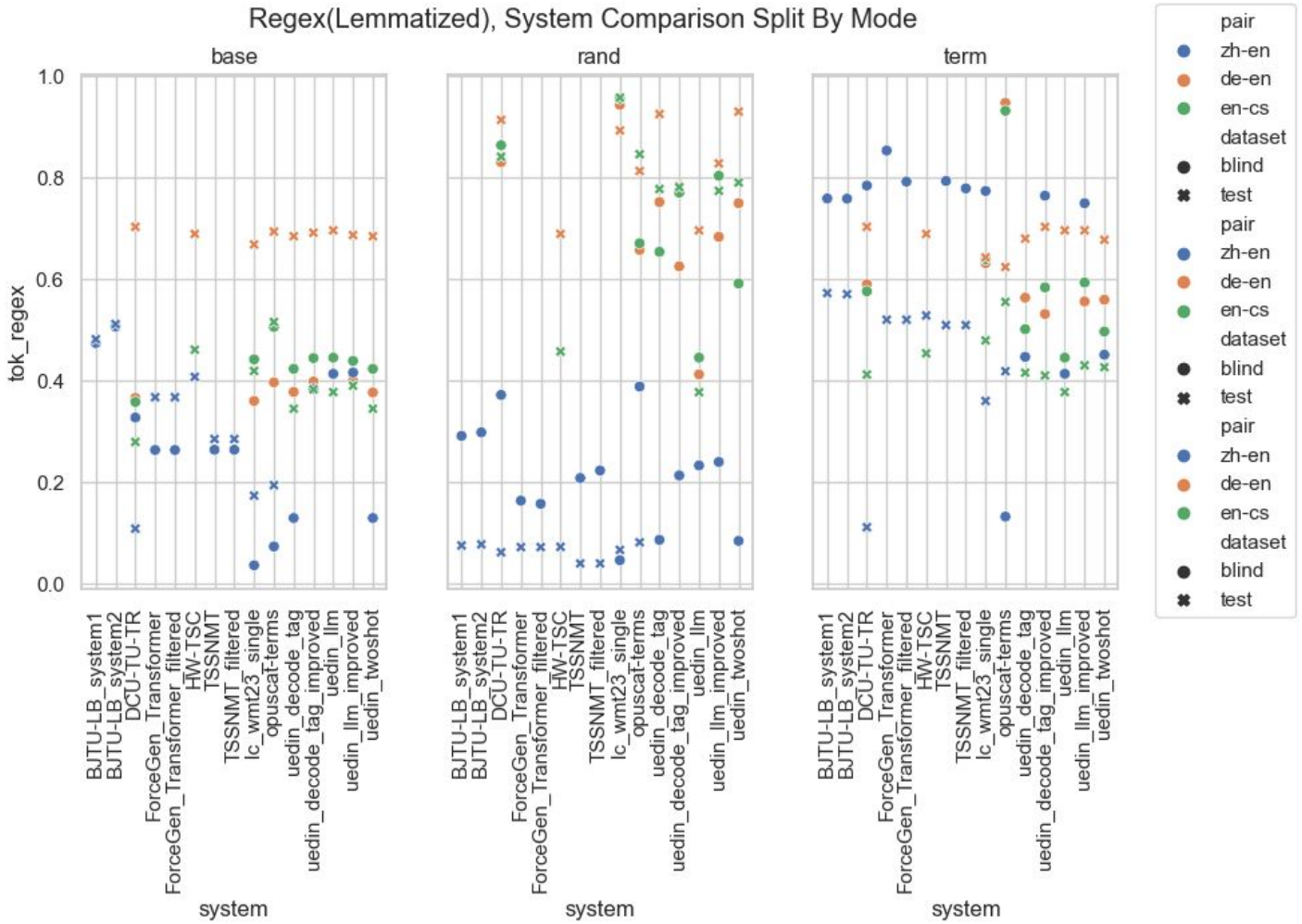

chrF, System Comparison Split By Mode (Blind Dataset Only)
base
rand
term
chrf
70
60
50
40
30
20
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
zh-en
de-en
en-cs

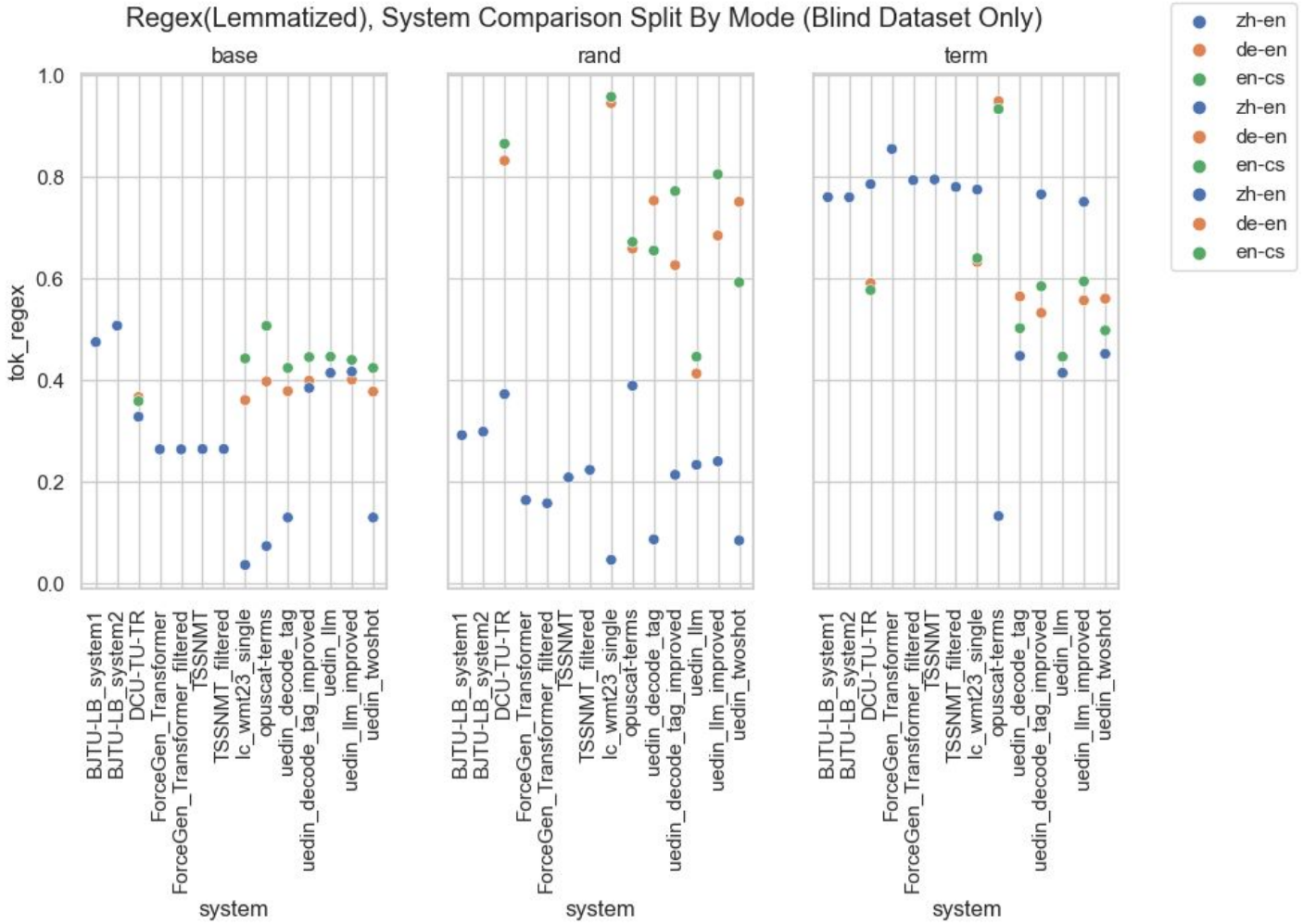

COMET-22-DA, System Comparison Split By Mode
base
rand
term
comet
0.9
0.8
0.7
0.6
0.5
0.4
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
HW-TSC
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
pair
zh-en
de-en
en-cs
dataset
blind
test

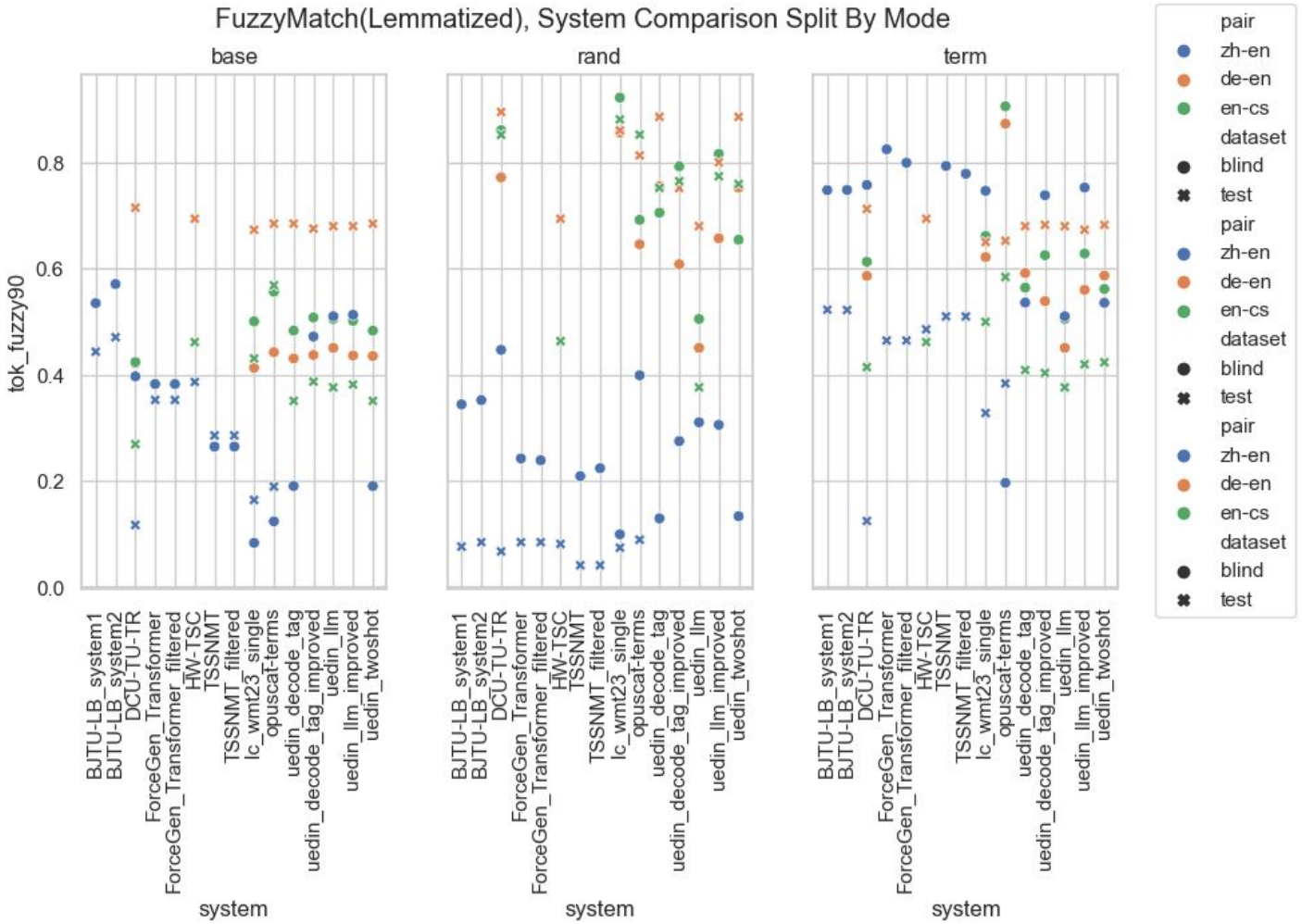

COMET-22-DA, System Comparison Split By Mode (Blind Dataset Only)
base
rand
term
comet
0.9
0.8
0.7
0.6
0.5
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
zh-en
de-en
en-cs

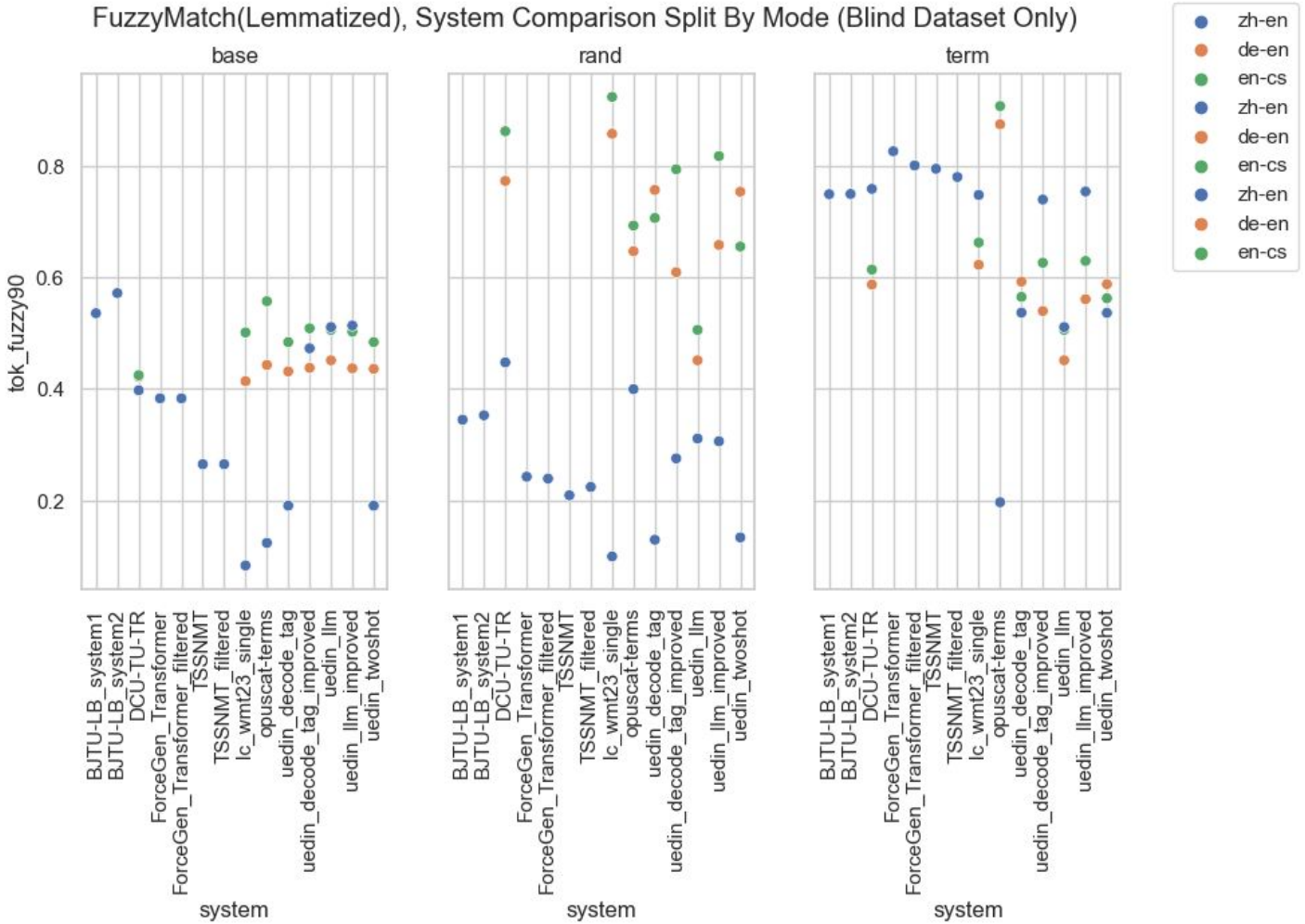

WMT22 Consistency Metric, System Comparison Split By Mode
base
rand
term
consistency
1.0
0.8
0.6
0.4
0.2
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
HW-TSC
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
pair
zh-en
de-en
en-cs
dataset
blind
test


WMT22 Consistency Metric, System Comparison Split By Mode (Blind Dataset Only)
base
rand
term
consistency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
zh-en
de-en
en-cs


Regex(Lemmatized), System Comparison Split By Mode
base
rand
term
tok_regex
1.0
0.8
0.6
0.4
0.2
0.0
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
HW-TSC
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
pair
zh-en
de-en
en-cs
dataset
blind
test

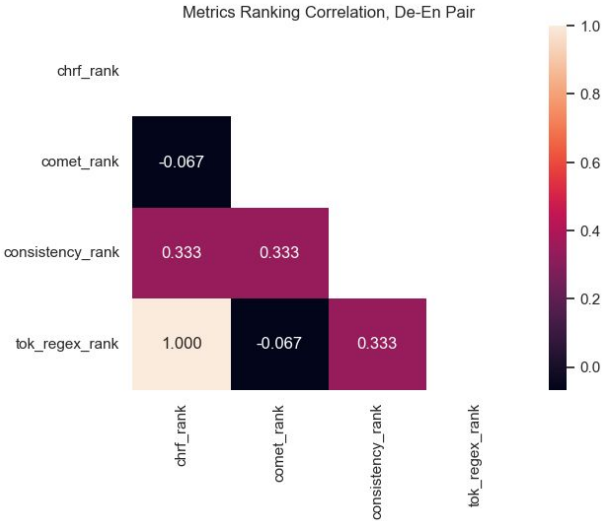

Regex(Lemmatized), System Comparison Split By Mode (Blind Dataset Only)
base
rand
term
tok_regex
1.0
0.8
0.6
0.4
0.2
0.0
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
zh-en
de-en
en-cs

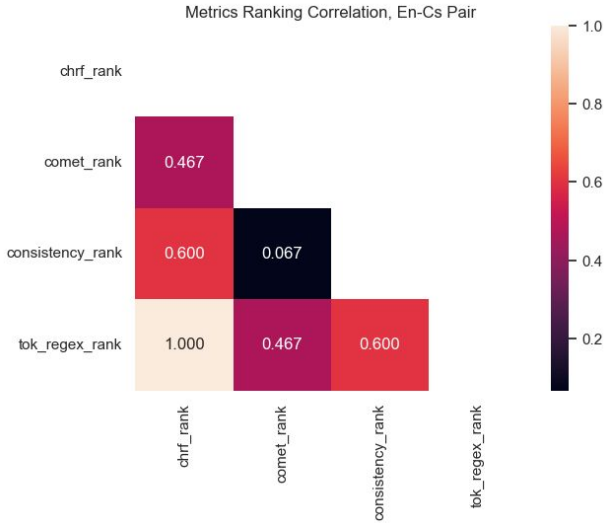

FuzzyMatch(Lemmatized), System Comparison Split By Mode
base
rand
term
tok_fuzzy90
0.0
0.2
0.4
0.6
0.8
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
HW-TSC
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
pair
zh-en
de-en
en-cs
dataset
blind
test

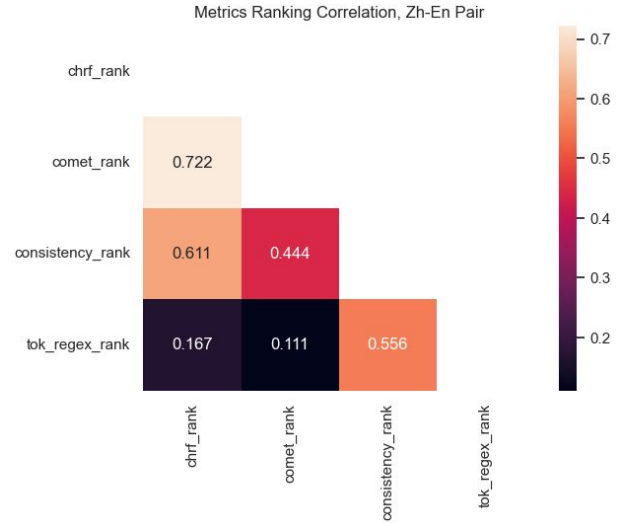

FuzzyMatch(Lemmatized), System Comparison Split By Mode (Blind Dataset Only)
base
rand
term
tok_fuzzy90
0.8
0.6
0.4
0.2
BJTU-LB_system1
BJTU-LB_system2
DCU-TU-TR
ForceGen_Transformer
ForceGen_Transformer_filtered
TSSNMT
TSSNMT_filtered
lc_wmt23_single
opuscat-terms
uedin_decode_tag
uedin_decode_tag_improved
uedin_llm
uedin_llm_improved
uedin_twoshot
system
zh-en
de-en
en-cs
zh-en
de-en
en-cs
zh-en
de-en
en-cs

Ranking Correlation: Main Metrics

The graphs represent Kenndall's tau measuring the correlation between the rankings by different metrics (higher is better).
NB: "tok_regex" means lemmatized regex metric

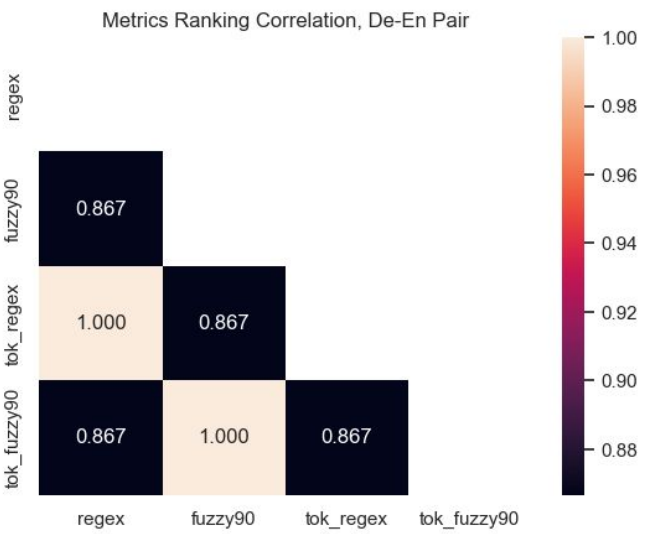


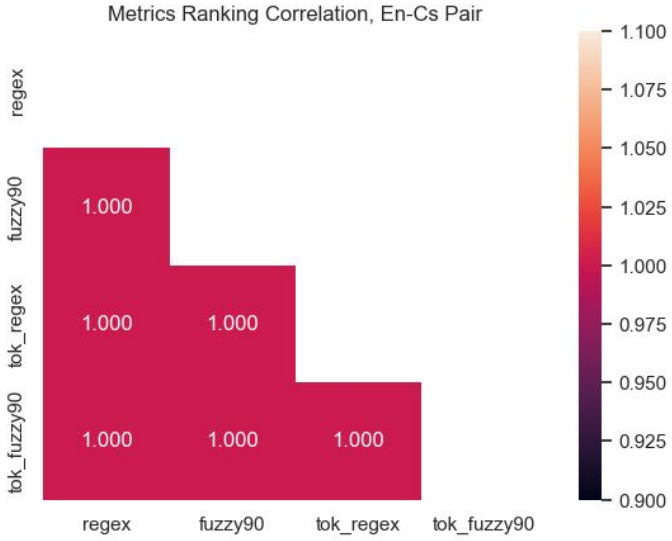


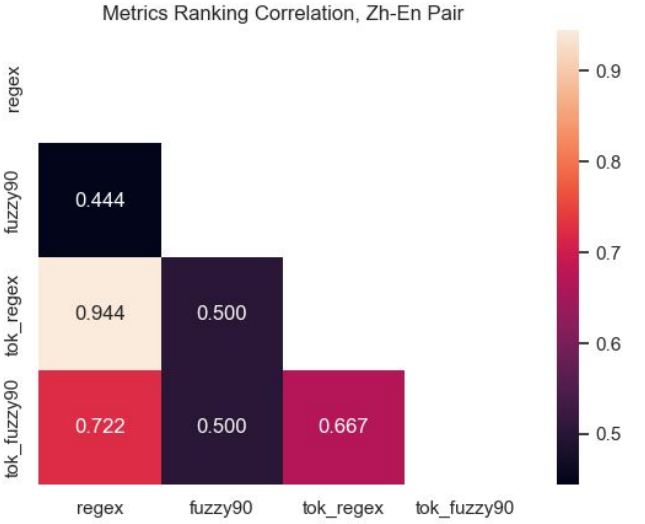


Ranking Correlation: Success Rate Variants

The graphs represent Kenndall's tau measuring the correlation between the rankings by different variants of success rate metrics (raw text VS lemmas, regex VS fuzzy match)
NB: "tok_" prefix means lemmatized metric







References

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-Centric Evaluation of Document-level Machine Translation with a New Densely Annotated Parallel Corpus of Novels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Rudolf Rosa and Vilém Zouhar. 2022. Czech and English abstracts of ÚFAL papers (2022-11-11). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

MuchMore Springer Bilingual Corpus. https://muchmore.dfki.de/resources1.htm

Kirill Semenov and Ondˇrej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.