Outline

DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation

Niccolò Campolungo*
Sapienza University of Rome
campolungo@di.uniroma1.it

Federico Martelli*
Sapienza University of Rome
martelli@di.uniroma1.it

Francesco Saina
SSML Carlo Bo, Rome
f.saina@ssmlcarlobo.it

Roberto Navigli
Sapienza University of Rome
navigli@diag.uniroma1.it

RAW-C: Relatedness of Ambiguous Words—in Context (A New Lexical Resource for English)

Sean Trott
University of California, San Diego
sttrott@ucsd.edu

Benjamin Bergen
University of California, San Diego
bkbergen@ucsd.edu

DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation

Niccolò Campolungo*
Sapienza University of Rome
campolungo@di.uniroma1.it

Federico Martelli*
Sapienza University of Rome
martelli@di.uniroma1.it

Francesco Saina
SSML Carlo Bo, Rome
f.saina@ssmlcarlobo.it

Roberto Navigli
Sapienza University of Rome
navigli@diag.uniroma1.it

Published on ACL'22 (avg review score of **4.375**)

Best Resource Paper

汇报人：刘柱

时　间：2022-05-15

“Ambiguation” in ACL’22

Sapienza NLP

Title	ACL	Type
DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation	Main (Long)	WSD Translation
ExtEnD: Extractive Entity Disambiguation	Main (Long)	Entity
Nibbling at the Hard Core of Word Sense Disambiguation	Main (Long)	WSD
Investigating Failures of Automatic Translation in the Case of Unambiguous Gender	Main (Long)	Translation
Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting	Main (Long)	WSD
Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem	Findings (Long)	Entity

Background

- The cause of incorrect translations: semantic biases?
- WSD biases: certain words towards more frequent meanings

at face value: as true or genuine without being questioned or doubted (Merriam-Webster.com)

WSD&MT

- WSD: to identify the meaning of a target word in a context.
- Target words: polysemous and homonymous words

  metaphor, entity, referential words… (more general)
- MFS bias: select the **m**ost **f**requent candidate **s**ense in the training data

  The words (form) with certain semantics (latent) are long-tailed distributed.

  It is rational to hypothesize that MT should “know” the different meanings
- Benchmarks have found some correlation between translation errors and MFS bias.

[Emelin et al. (2020)]

Related works

Benchmarks:

- ContraWSD (DE↔EN)
- Another WSD Test Suit (DE→EN, translation only covers one sense)
- MuCoW (16 language pairs, more than 200K pairs, automatically)

Some investigating disambiguation capabilities by…

- Exploring internal and contextual representations
- A statistical method for the correlation of sense and translation error

Drawbacks

- Not based on entirely manually-curated benchmarks
- Rely heavily on automatically generated resources to determine the correctness of a translation
- Not cover multiple language combinations

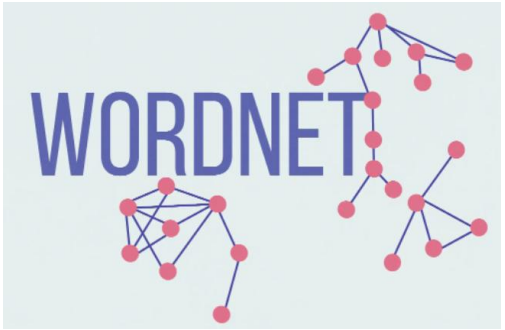
Contributions

- DIBIMT: gold-quality test bed with five languages.
- Four novel metrics to better clarify the semantic biases in MT models.
- A thorough statistical and linguistic analysis for 7 SOTA MT systems.

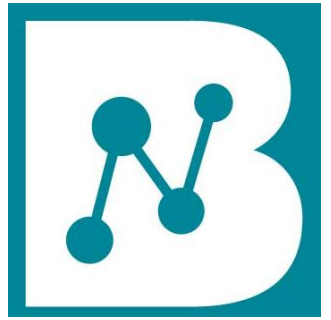
Building DIBIMT

- Source (EN): context with a polysemous target words.
- Target: corresponding good and bad lexicalization sets in a specific language.

  (DE, IT, ES, RU, ZH)

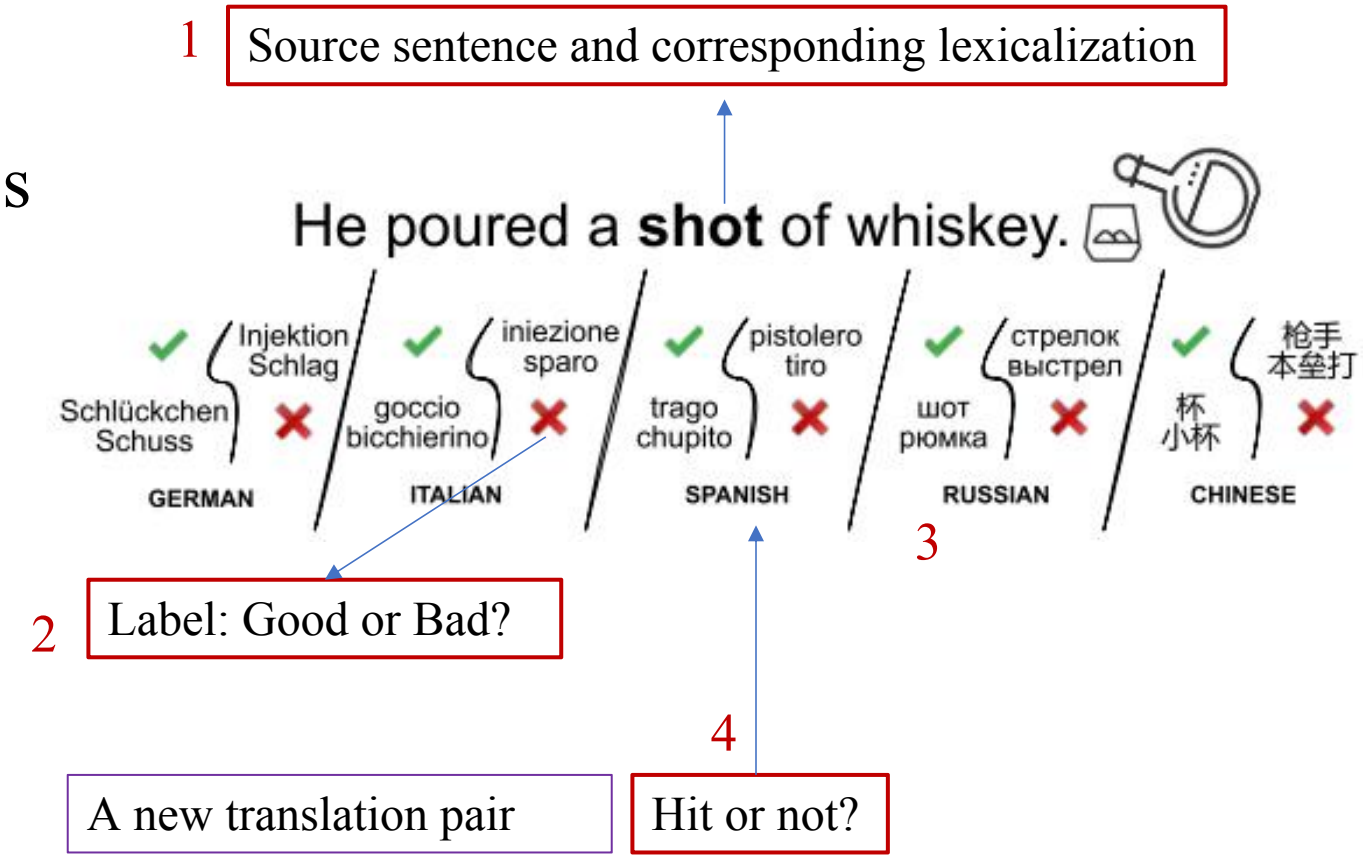


Wiktionary
The free dictionary



Building DIBIMT

1. Sentence Selection Process
2. Annotation
3. Resulting Dataset
4. Analysis Procedure



Building DIBIMT

1. Sentence Selection Process
2. Annotation
3. Resulting Dataset
4. Analysis Procedure

Aim: *initial items* - $X = (s, w_i, \sigma)$

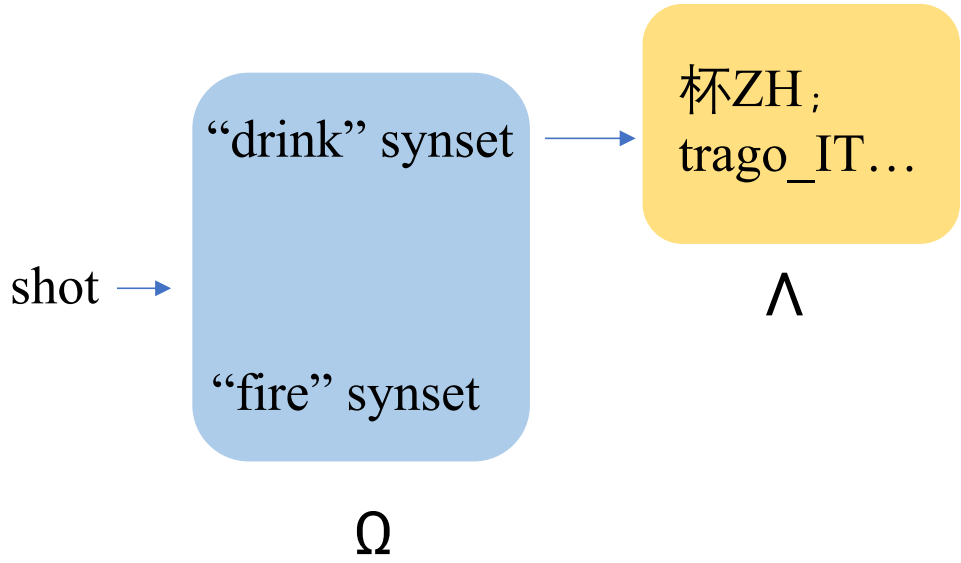
- Starting Sentence Pool
  - Wordnet
  - Wiktionary: additional synset mapping
  - Babelnet: synset σ
- Sentence Filtering
  - Polysemous:
  - Light overhead: one sentence per sense per source.
  - Target synsets do not have to be monosemous.

Building DIBIMT

1. Sentence Selection Process
2. Annotation
3. Resulting Dataset
4. Analysis Procedure

Aim: *annotated items* - $X = (s, w_i, \sigma, G_L, B_L)$

- Pre-annotation Item Creation
  - $\mathcal{G}_L = \Lambda_L(\sigma)$
  - $\mathcal{B}_L = \bigcup_{\hat{\sigma} \in \Omega_L(\lambda_P^x) \setminus \{\sigma\}} \Lambda_L(\hat{\sigma})$



Building DIBIMT

1. Sentence Selection Process
2. Annotation
3. Resulting Dataset
4. Analysis Procedure

Aim: *annotated items* - $X = (s, w_i, \sigma, G_L, B_L)$

- Pre-annotation Item Creation
  - $\mathcal{G}_L = \Lambda_L(\sigma)$
  - $\mathcal{B}_L = \bigcup_{\hat{\sigma} \in \Omega_L(\lambda_P^x) \setminus \{\sigma\}} \Lambda_L(\hat{\sigma})$

- Annotation Guidelines
  - To update G_L and B_L by adding new good translations or removing improper ones (from G_L to B_L)
  - To discard sentences (idiomatic expressions, proper noun or insufficient to disambiguate target words)

Building DIBIMT

1. Sentence Selection Process
2. Annotation
3. Resulting Dataset
4. Analysis Procedure

	All	Nouns	Verbs
# items	597	314	286
# lemmas	305	186	147
# synsets	471	254	217

Table 1: General statistics of our annotated dataset. POS-specific lemmas do not sum to "All" as they can overlap across POS tags (e.g., run).

Building DIBIMT

1. Sentence Selection Process
2. Annotation
3. Resulting Dataset
4. Analysis Procedure

	%OG	%RG	%SL
DE	50.9	25.0	59.7
ES	49.6	19.5	47.7
IT	49.1	38.2	67.1
RU	67.4	57.3	54.4
ZH	55.2	69.0	46.3
Mean	54.4	41.8	55.0

Table 2: Annotation Statistics: %OG represents the average percentage of Good lemmas that are Original, i.e., were added by our annotators; %RG represents the average percentage of Good lemmas that were Removed, i.e., lemmas that came from BabelNet and that our annotators deemed incorrect in the context of the given example; %SL represents the average percentage of times two senses Share the same set of Lexicalizations for two different example sentences.

Low SL: the translation of synonyms is unlikely to be synonym (Context is all)

Building DIBIMT

1. Sentence Selection Process
2. Annotation
3. Resulting Dataset
4. Analysis Procedure

Aim: *analyzed items* - $X_L^M = (X_L, t_L, R, w_L)$

- Translate, split and hit
  - $t_L = M_L(s)$ # translated sentence
  - Tokenize, POS, lemmatize of t_L
  - Any hit:

    $[R]$: GOOD (G_L), BAD (B_L) and MISS

    $[w_L]$: hit word

Experiments

- Comparison systems

  DeepL Translator： a SOTA commercial NMT system

  Google Translate: popular commercial system

  OPUS: smallest (74M) SOTA

  MBart50: 50 languages (610M)

  M2M100: 100 languages (418M & 1.2B)

DeepL

https://www.deepl.com/translator

DeepL翻译

線上翻譯服務

语言　　下载PDF　监视　编辑

DeepL翻译（英语：DeepL Translator）是2017年8月由总部位于德国科隆的DeepL GmbH（一家由Linguee支持的创业公司）推出的免费神经机器翻译服务[2][3][4]。评论家对于它的评价普遍正面，认为它的翻译比起Google翻译更为准确自然[5][6]。

DeepL目前支援保加利亚语、简体中文、捷克语、丹麦语、荷兰语、英语（美式或英式）、爱沙尼亚语、芬兰语、法语、德语、希腊语、匈牙利语、意大利语、日语、拉脱维亚语、立陶宛语、波兰语、葡萄牙语、巴西葡萄牙语、罗马尼亚语、俄语、斯洛伐克语、斯洛文尼亚语、西班牙语、瑞典语之间的翻译。总计24种语言，552种语言组合[4]。除此之外，DeepL在翻译上述语言时采用了语言等效性原理，其会先在后台翻译作英语，然后再翻译成另一种语言。

DeepL并不会在官网上显示任何广告。开发公司希望借由出售API授权来盈利[7]。

DeepL翻译
DeepL Translator

DeepL

网站类型	机器翻译

DeepL

评价

对DeepL Translator的反应普遍是正面的，TechCrunch对其翻译的准确性表示赞赏，称其比Google翻译更准确、更细致[5]，《世界报》则感谢其开发者将法语文本翻译成"听起来更像法语"的表达方式[34]。荷兰RTL Z电视频道网站的一篇新闻文章称，DeepL Translator"在荷兰语到英语的过程中提供了更好的翻译[……]，反之亦然"[35]。

其他新闻机构和新闻网站也对DeepL Translator表示了赞许，如意大利报纸《共和报（英语：la Repubblica）》[36]和拉美网站WWWWhat's new?[37]。

发布时，DeepL宣称在盲测中超越了其他竞争对手，包括谷歌翻译、亚马逊翻译、微软翻译和Facebook[38][39][40][41][42][43][44]，但尚未有独立测评对这些服务进行比较[45]。具体来说，每个服务都提取了119段不同领域的文本并进行翻译，然后由公司外部的专家翻译对译文进行评估[18]。此外，即便是要翻译的日文混杂着方言，其结果依旧准确[46]。DeepL Translator获得了2020年Webby最佳实践奖，以及2020年Webby技术成就奖（应用、移动和功能），均为应用、移动和语音类[47]。

Results

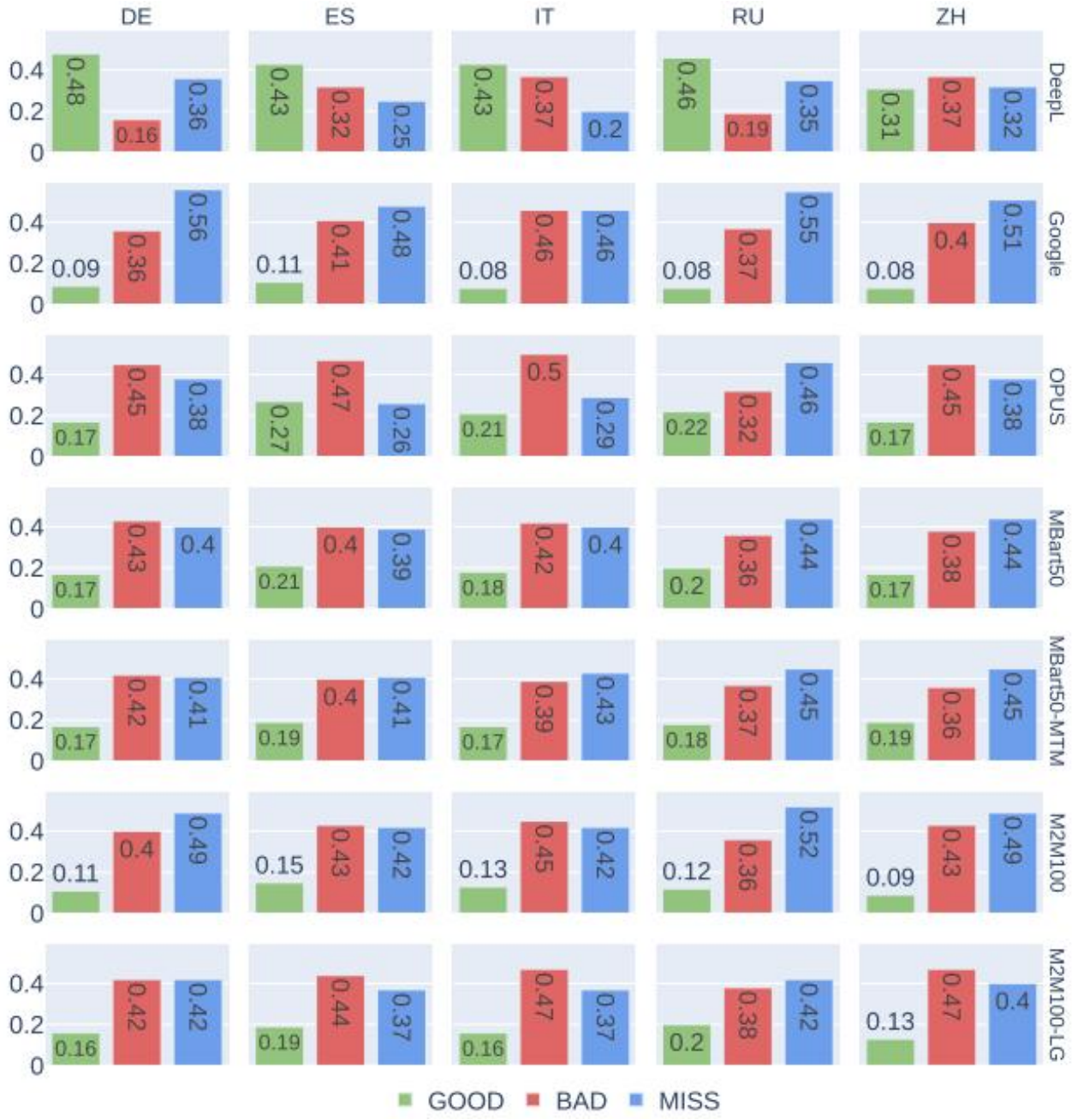
- Discussion of MISS
- Reason (human inspection on random 70 samples):

1) ~19% word omission (ZH ES)

2) ~11% tokenization (ZH RU)

3) ~5% self-translations

4) ~23% nothing to do with source text

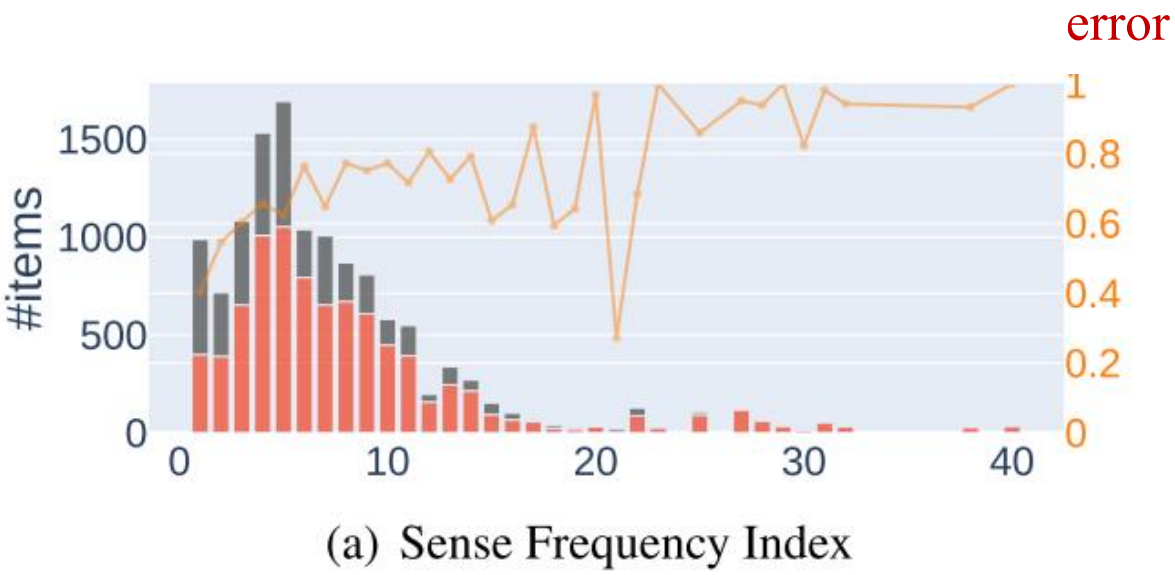


General results

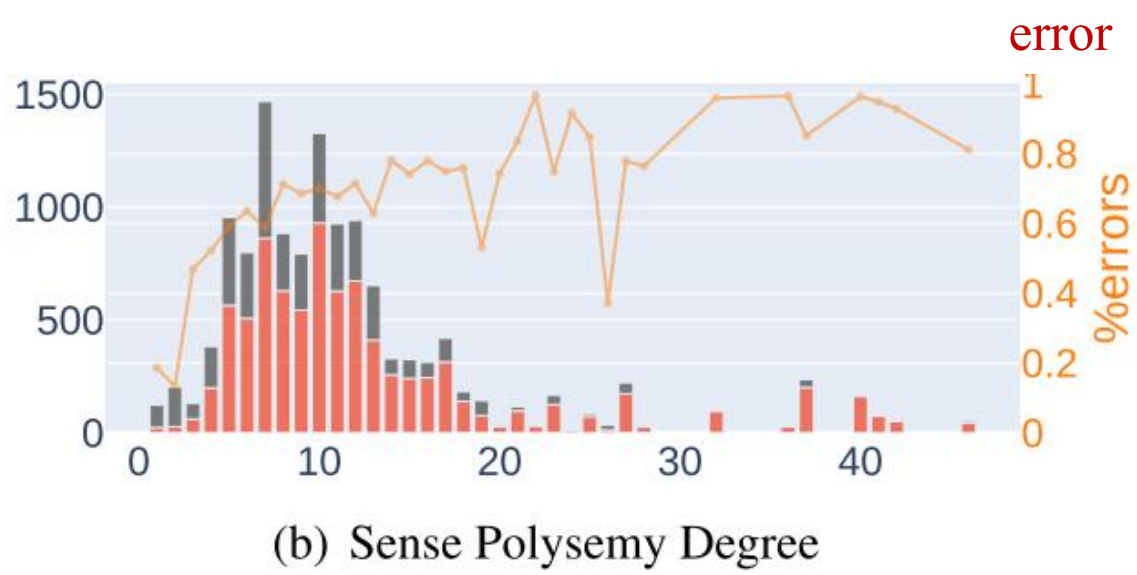
	DeepL	Google	M2M100	$M2M100_{LG}$	MBart50	$MBart50_{MTM}$	OPUS	Mean
DE	74.60	21.90	22.19	26.96	28.73	28.65	27.99	33.00
ES	57.87	22.54	25.51	30.00	33.89	32.66	36.66	34.16
IT	53.49	18.04	21.83	25.14	29.34	30.54	29.95	29.76
RU	71.58	22.89	26.22	35.19	36.06	33.33	41.07	38.05
ZH	46.00	15.04	16.99	22.35	31.21	34.15	27.75	27.64
Mean	60.71	20.08	22.55	27.93	31.85	31.87	32.68	32.52

- Acc = #GOOD / (#GOOD + #BAD)

Semantic Biases

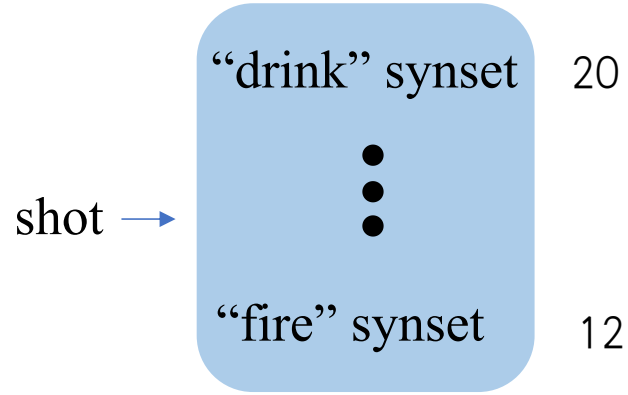


(a) Sense Frequency Index



(b) Sense Polysemy Degree

- SFII (Sense Frequency Index Influence): index of the target synset σ in Λ ordered according to frequency in WordNet
- SPDI (Sense Polysemy Degree Importance): degree of polysemy
- Less frequent and more senses will cause more errors.



Semantic Biases

	DeepL		Google		M2M100		$M2M100_{LG}$		MBart50		$MBart50_{MTM}$		OPUS		Mean	
	SFII	SPDI	SFII	SPDI	SFII	SPDI	SFII	SPDI	SFII	SPDI	SFII	SPDI	SFII	SPDI	SFII	SPDI
DE	34.78	28.30	86.61	79.54	82.00	76.15	78.90	74.71	84.10	73.86	84.95	74.24	79.85	76.25	75.89	69.00
ES	56.04	46.14	83.84	78.41	83.08	77.95	79.87	73.84	77.13	71.06	79.06	71.57	74.85	69.12	76.27	69.73
IT	57.71	49.01	85.47	80.62	80.22	76.58	78.69	76.10	78.67	71.51	79.41	69.48	80.59	72.02	77.25	70.76
RU	41.97	33.64	84.01	83.49	79.85	78.34	74.72	69.69	73.86	70.11	78.58	72.87	68.49	69.27	71.64	68.20
ZH	64.97	59.58	91.97	87.98	91.81	87.18	88.79	82.17	80.39	73.14	76.59	71.50	79.96	75.66	82.07	76.75
Mean	51.10	43.33	86.38	82.01	83.39	79.24	80.19	75.30	78.83	71.94	79.72	71.93	76.75	72.46	76.62	70.89

- Average error for SFII group and SPDI group.

Semantic Biases

	DeepL		Google		M2M100		$M2M100_{LG}$		MBart50		$MBart50_{MTM}$		OPUS		Mean	
	MFS	MFS+	MFS	MFS+	MFS	MFS+	MFS	MFS+	MFS	MFS+	MFS	MFS+	MFS	MFS+	MFS	MFS+
DE	53.68	84.21	56.76	86.82	61.28	87.23	59.13	87.30	58.89	89.72	55.82	89.56	56.98	87.92	57.51	87.54
ES	59.89	87.91	61.96	89.05	61.81	89.37	61.78	88.03	60.17	91.10	63.09	91.85	64.47	91.21	61.88	89.79
IT	68.08	86.38	61.96	87.23	60.75	86.79	62.82	88.81	62.90	87.50	68.97	91.81	64.48	89.66	64.28	88.31
RU	50.00	83.33	48.12	83.28	47.87	83.41	45.25	84.16	47.39	87.20	44.91	87.96	48.40	84.04	47.42	84.77
ZH	49.07	88.89	56.05	88.20	59.06	91.34	59.35	92.45	50.66	89.87	54.17	90.28	51.71	87.45	54.30	89.78
Mean	56.14	86.15	56.97	86.92	58.15	87.63	57.66	88.15	56.00	89.08	57.39	90.29	57.21	88.06	57.08	88.04

- MFS+: The frequency that the SFII of a BAD translation is lower (more frequent) than that of the target words.
- MFS: the BAD translation is the most frequent.

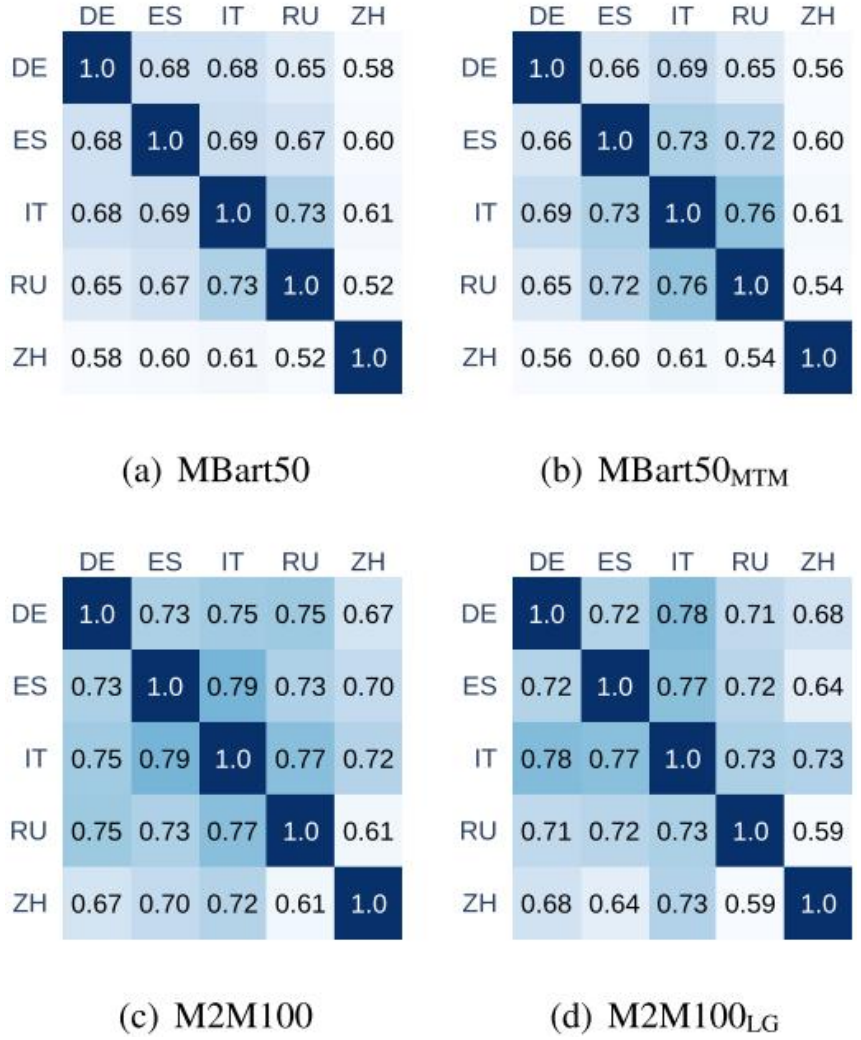
Are verbs harder than nouns?

	ALL	NOUN	VERB
Accuracy	32.11	34.15	30.02
%MISS	38.03	29.36	47.57
MFS	57.86	60.13	52.60
MFS+	88.68	87.57	88.74
SFII	76.98	69.16	76.90
SPDI	70.80	66.86	72.87

- In WSD, verbs are generally harder than nouns due to their highly polysemous nature [Barba et al. 2021]

Is the encoder disambiguating?

- If the encoder is the sole contributor to disambiguating, the meaning is always the same regardless of the target language.
- How often do L1 and L2 have a synset in common?
- ~70%
- ZH is not as compatible.



How challenging is DIBIMT?

- Above is ESCHER's WSD accuracy (80.7 on WSD ALL v. 67.84 here).
- NMT models are still not on par with dedicated WSD systems.

	DeepL	Google	M2M	$M2M_{LG}$	MB	MB_{MTM}	OPUS	Mean
DE	66.86	71.04	65.85	67.18	66.87	67.77	66.95	67.50
ES	67.89	72.76	66.77	66.86	65.37	67.18	66.83	67.67
IT	66.67	72.58	66.35	68.50	64.33	65.81	65.82	67.15
RU	66.76	69.55	66.42	67.69	66.35	64.29	69.21	67.18
ZH	68.42	71.89	69.26	69.82	68.93	69.58	69.88	69.68
Mean	67.32	71.56	66.93	68.01	66.37	66.93	67.74	67.84

	DeepL	Google	M2M100	$M2M100_{LG}$	MBart50	$MBart50_{MTM}$	OPUS	Mean
DE	74.60	21.90	22.19	26.96	28.73	28.65	27.99	33.00
ES	57.87	22.54	25.51	30.00	33.89	32.66	36.66	34.16
IT	53.49	18.04	21.83	25.14	29.34	30.54	29.95	29.76
RU	71.58	22.89	26.22	35.19	36.06	33.33	41.07	38.05
ZH	46.00	15.04	16.99	22.35	31.21	34.15	27.75	27.64
Mean	60.71	20.08	22.55	27.93	31.85	31.87	32.68	32.52

Is this a decoding issue?

- Model Errors: percentage of times a model thought its BAD translation was better than a GOOD one
- Sampling 50 times, how often the perplexities meet $p_{BAD} > p_{GOOD}$?
- Most semantic biases are not caused by the decoding strategy.

	M2M100	$\text{M2M100}_{\text{LG}}$	MBart50	$\text{MBart50}_{\text{MTM}}$	OPUS	Mean
DE	98.00	98.00	92.00	94.00	84.00	93.20
ES	100.00	98.00	88.00	90.00	94.00	94.00
IT	94.00	90.00	86.00	100.00	88.00	91.60
RU	94.00	90.00	98.00	92.00	88.00	92.40
ZH	96.00	98.00	94.00	98.00	92.00	95.60
Mean	96.40	94.80	91.60	94.80	89.20	93.36

Conclusions

- DIBIMT (5 languages, 7 systems) for measuring and understanding semantic biases in NMT.
- Founding: synsets' lexicalizations cannot be interchangeable.
- Future: high MISS cases; widening language coverage and increasing number of sentences.

RAW-C: Relatedness of Ambiguous Words—in Context (A New Lexical Resource for English)

Sean Trott
University of California, San Diego
sttrott@ucsd.edu

Benjamin Bergen
University of California, San Diego
bkbergen@ucsd.edu

Published on ACL'21

Motivation

- 7% homonymous; 84% polysemous
- Graded/continuous nature: “the boy runs” v. “the cheetah runs”. (different mental images)
- Dynamic and context-dependent
- Metrics (model) should meet:
  - Criterion 1: Disambiguation
  - Criterion 2: Contextual Gradation

  WSD only satisfies the first Criterion

- RAW-C: **R**elatedness of **A**mbiguous **W**ords – in **C**ontext.

Related work

- De-contextualized Word Similarity and Relatedness

Isolated; for static semantic representations (e.g., GloVe); Only C2

- WSD

Only C1; how meaning modulated within a given sense category? (e.g. run)

- Contextualized Word Similarity and Relatedness (C2)

SCWS (12% the same wordform); WiC (binary classification); CoSimLex

- Contextualized Similarity of Ambiguous Words

One dataset [Haber et al, 2020] C1 & C2, but size limitation and without context control

Contribution

- A dataset satisfying C1 (Disambiguation) and C2 (Contextual Gradation)
- With tightly controlled sentence contexts: inflection and POS.

1a. He saw a *fruit* bat.
1b. He saw a *furry* bat.
2a. He saw a *wooden* bat.
2b. He saw a *baseball* bat.

Dataset building

- 112 target words inspired by past psycholinguistic studies
- For each word, four sentence (six pairs) are constructed.

1a. He saw a *fruit* bat.
1b. He saw a *furry* bat.
2a. He saw a *wooden* bat.
2b. He saw a *baseball* bat.

H or P?

bat	He saw a furry bat.	He saw a wooden bat.	FALSE	Homonymy
bat	He saw a furry bat.	He saw a baseball bat.	FALSE	Homonymy
bat	He saw a fruit bat.	He saw a wooden bat.	FALSE	Homonymy
bat	He saw a fruit bat.	He saw a baseball bat.	FALSE	Homonymy
bat	He saw a furry bat.	He saw a fruit bat.	TRUE	Homonymy
bat	He saw a wooden bat.	He saw a baseball bat.	TRUE	Homonymy

Same-Sense type

Same-Sense type: True (1a-1b & 2a-2b) or False (check in a dictionary)
Homonymy or Polysemy?
According to distinct meaning 1 and 2, refer to Merriam-Webster Dictionary and OED.
NOTE that it is word-specific ***not*** semantics-specific.

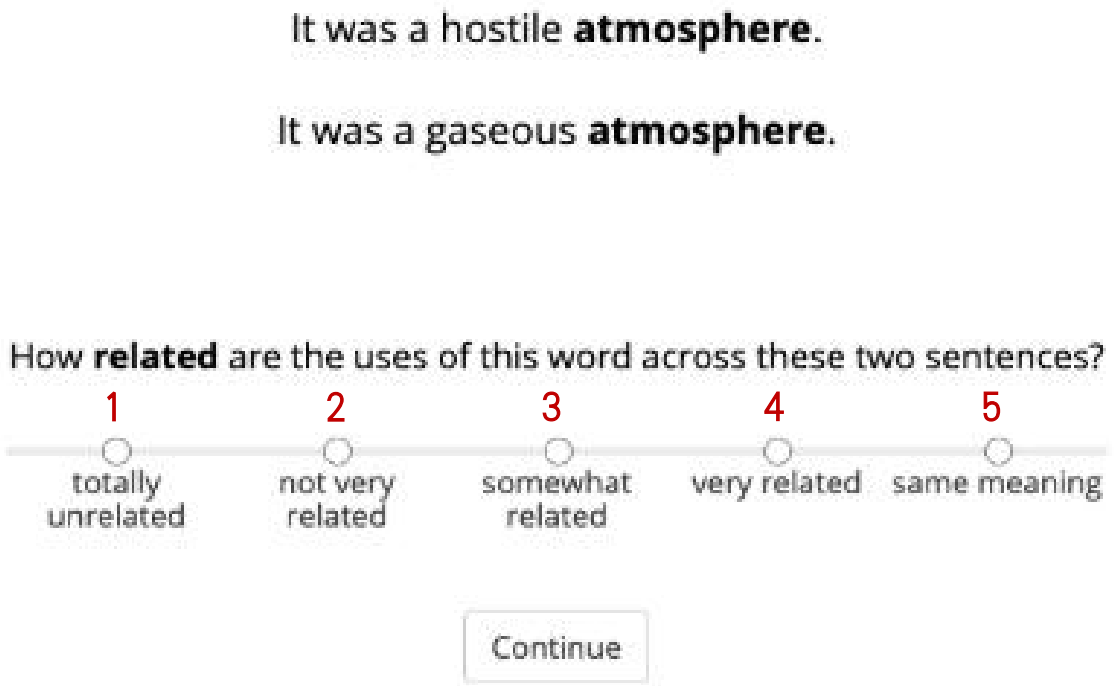
Statistics

Ambiguity Type	#Words	#Sentence Pairs
Homonymy	38	228
Polysemy	74	444

NOUN	VERB	ALL
84	28	112

Human Annotation

- Score of relatedness for each pair.
- 77 participants to label individually.
- 115 * 12 = 1380 items



Inter-Annotator Agreement

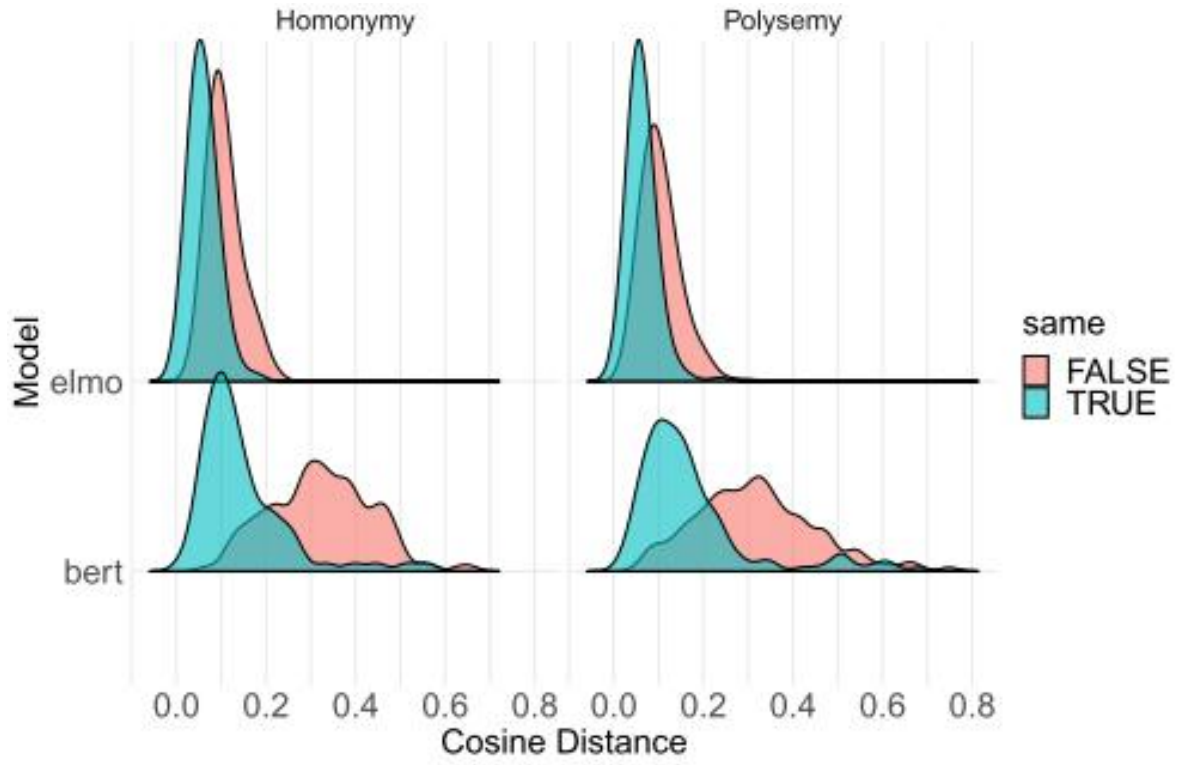
- To what extent do each participant's responses correlate with the consensus rating by the 76 other participants?
- One v. Mean Relatedness of 76
- Spearman's r: ρ
- Avg: 0.79

Analysis of Sentence Pairs

- Representation distributions of COS distance between target words by elmo and bert.

[Looks like bert can distinguish two types]

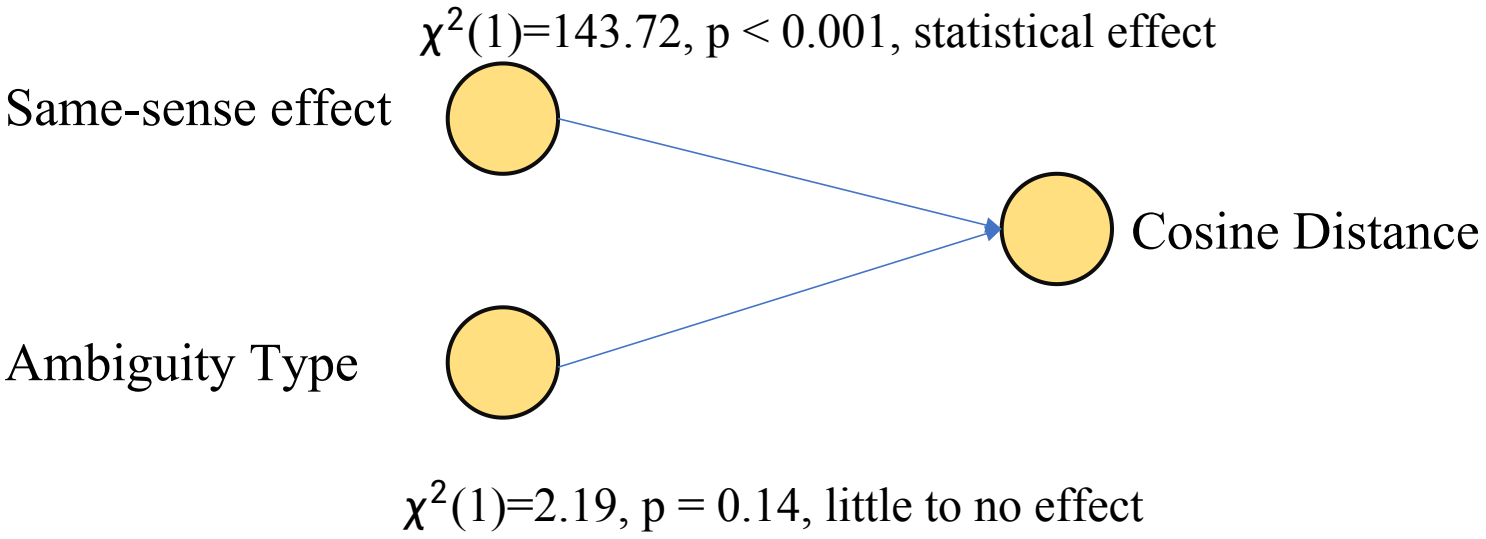
- Two types
- [Confusion] What's the differences of the SAME-sense in homo. and poly.?



bat	He saw a wooden bat.	He saw a baseball bat.	TRUE	Homonymy

More statistical analyses

- Log-likelihood ratio test using lme4 package in R



Is it fair to claim that “their ability to discriminate between Homonymy and Polysemy was marginal at best”?

Analysis of Human Annotations

- Same-Sense type

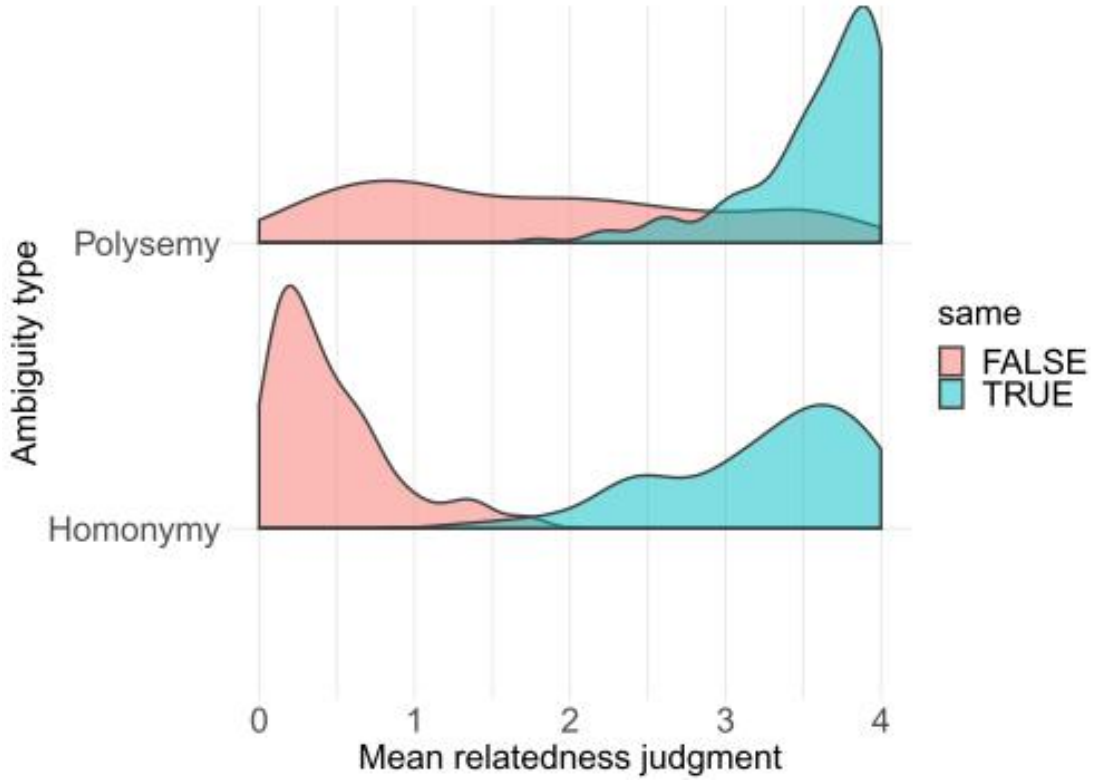
  p<.001; median 4 (T) v. 1 (F)
- Homo v. Poly type

  p<.001
- Can Cosine Distance explain independent variance?

  BERT: [$\chi^2(1)=36.19$, p<.001]

  ELMo: [$\chi^2(1)=16.92$, p<.001]

  BERT looks better.



Evaluation of Language Models

- How Cosine distance relates to the human Relatedness?

Pearson's r: 0.58 (Bert) 0.53 (Elmo) v. 0.79 (inter-annotator)

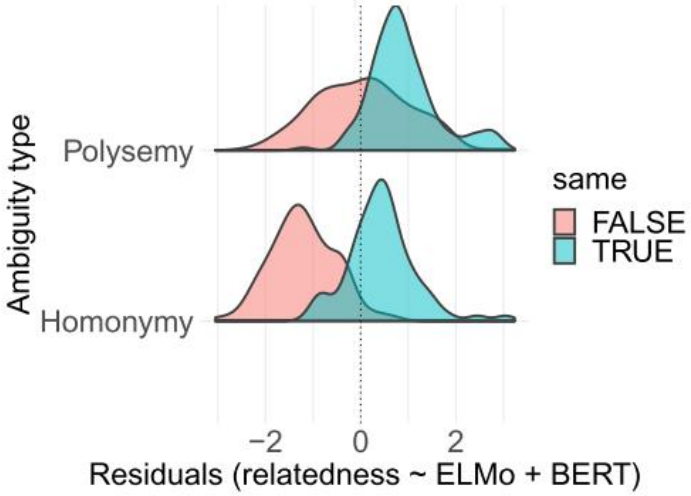
- Linear regression between these two terms:

1) R^2 (explain the fitness of the curve) 0.37 (v. 0.66 +Two Types; 0.71 All)

2) Residual analysis:

Underestimate in Same-sense (both Poly & Homo)

Overestimate in Different-sense in Homo



Conclusion&Discussion

- RAW-C: relatedness judgement; tightly controlled context; Two Types
- Primary findings:

1) Representations from both models capture Same-Sense type uses, but marginally discriminate Ambiguous type.
2) Distances from both models explain variance in human Relatedness.
3) Given two Types, both models have overestimation and underestimation.

Future Work

- Meaning or form? [Bender and Koller, 2020] Human fulfill the Disambiguation Criterion and the Contextual Gradation Criterion.
- Look to mental lexicon: continuity v. categorical
- How to incorporate high-level sense knowledge into internal representations? (SenseBERT)

Thanks

A.1 Best papers in ACL'22

- Best Paper
  - **Learned Incremental Representations for Parsing** (Nikita Kitaev, Thomas Lu and Dan Klein)
- Best Special Theme Paper
  - **Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization** (Aidan Pine, Dan Wells, Nathan Brinklow, Patrick William Littell and Korin Richmond)
- Best Resource Paper
  - **DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation** (Niccolò Campolungo, Federico Martelli, Francesco Saina and Roberto Navigli)
- Best Linguistic Insight Paper
  - **KinyaBERT: a Morphology-aware Kinyarwanda Language Model** (Antoine Nzeyimana and Andre Niyongabo Rubungo)

https://www.2022.aclweb.org/best-paper-awards

A.1 Best papers in ACL'22

- Outstanding Papers
  - **Evaluating Factuality in Text Simplification** (By Ashwin Devaraj, William Berkeley Sheffield, Byron C Wallace and Junyi Jessy Li)
  - **Online Semantic Parsing for Latency Reduction in Task-Oriented Dialogue** (Jiawei Zhou, Jason Eisner, Michael Newman, Emmanouil Antonios Platanios and Sam Thomson)
  - **Learning to Generalize to More: Continuous Semantic Augmentation for Neural Machine Translation** (Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo and Rong Jin)
  - **Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity** (Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel and Pontus Stenetorp)
  - **Inducing Positive Perspectives with Text Reframing** (Caleb Ziems, Minzhi Li, Anthony Zhang and Diyi Yang)
  - **Ditch the Gold Standard: Re-evaluating Conversational Question Answering** (Huihan Li, Tianyu Gao, Manan Goenka and Danqi Chen)
  - **Active Evaluation: Efficient NLG Evaluation with Few Pairwise Comparisons** (Akash Kumar Mohankumar and Mitesh M Khapra)
  - **Compression of Generative Pre-trained Language Models via Quantization** (Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, Ngai Wong)

https://www.2022.aclweb.org/best-paper-awards