Nyelvi sokszínűség az emberi és gépi fordításban

Recski Gábor

TU Wien
gabor.recski@tuwien.ac.at

Elektrubadúr plusz, 2024.05.15

Almae Matres

2004–2010	ELTE Angol-Amerikai Intézet
2005–2017	MTA-ELTE Elméleti Nyelvészet Tanszék
2009–2013	MTA SZTAKI
2014–2016	MTA Nyelvtudományi Intézet
2016–2018	BME VIK
2020–	TU Wien

Tartalom

- Mit tanulnak a gépi fordítók?
- Nyelvi diverzitás
- Kísérleti eredmények

Mit tanulnak a gépi fordítók?

Mit tanulnak a gépi fordítók?

Olyan szöveget alkotni, ami hasonlít a tanítóadatra.

Mit tanulnak a gépi fordítók?

Olyan szöveget alkotni, ami hasonlít a tanítóadatra.

Ez több (megoldatlan) problémát is felvet.

Mit tanulnak a gépi fordítók?

Olyan szöveget alkotni, ami hasonlít a tanítóadatra.

Ez több (megoldatlan) problémát is felvet.

- Mi a (jó) tanítóadat?

Mit tanulnak a gépi fordítók?

Olyan szöveget alkotni, ami hasonlít a tanítóadatra.

Ez több (megoldatlan) problémát is felvet.

- Mi a (jó) tanítóadat?
- Mi az, hogy hasonlítani?

Mit tanulnak a gépi fordítók?

Olyan szöveget alkotni, ami hasonlít a tanítóadatra.

Ez több (megoldatlan) problémát is felvet.

- Mi a (jó) tanítóadat?
- Mi az, hogy hasonlítani?
- És mi lesz így a sokféleséggel?

Nyelvi diverzitás

- Lexikai diverzitás
- Grammatikai diverzitás

Néhány kísérleti eredmény

Vanmassenhove, E., Shterionov, D, and Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pp. 2203–2213

Néhány kísérleti eredmény

Vanmassenhove, E., Shterionov, D, and Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pp. 2203–2213

- 3 MT rendszer, 6 nyelvpár (EN→FR, EN→ES, FR→EN, ES→EN)

Néhány kísérleti eredmény

Vanmassenhove, E., Shterionov, D, and Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pp. 2203–2213

- 3 MT rendszer, 6 nyelvpár (EN→FR, EN→ES, FR→EN, ES→EN)
- lexikai és morfológiai (alaktani) diverzitást is mér, többféle metrikával

Néhány kísérleti eredmény

Vanmassenhove, E., Shterionov, D, and Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pp. 2203–2213

- 3 MT rendszer, 6 nyelvpár (EN→FR, EN→ES, FR→EN, ES→EN)
- lexikai és morfológiai (alaktani) diverzitást is mér, többféle metrikával
- Europarl korpusz (EU Parlament jegyzőkönyvei, Koehn 2005)

Néhány kísérleti eredmény

Vanmassenhove, E., Shterionov, D, and Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pp. 2203–2213

- 3 MT rendszer, 6 nyelvpár (EN→FR, EN→ES, FR→EN, ES→EN)
- lexikai és morfológiai (alaktani) diverzitást is mér, többféle metrikával
- Europarl korpusz (EU Parlament jegyzőkönyvei, Koehn 2005)
- **Az emberi fordítás lexikai és morfológiai diverzitása is nagyobb a gépinél**

Néhány kísérleti eredmény

Néhány kísérleti eredmény

Toral, A (2019). Post-editese: an Exacerbated Translationese. In *Proceedings of Machine Translation Summit XVII*, pp. 273–281

Néhány kísérleti eredmény

Toral, A (2019). Post-editese: an Exacerbated Translationese. In *Proceedings of Machine Translation Summit XVII*, pp. 273–281

- Az ember által javított gépi fordítást vizsgálja (Post-editing, PE)

Néhány kísérleti eredmény

Toral, A (2019). Post-editese: an Exacerbated Translationese. In *Proceedings of Machine Translation Summit XVII*, pp. 273–281

- Az ember által javított gépi fordítást vizsgálja (Post-editing, PE)
- Három adathalmaz (újsághírek, filmfeliratok), öt nyelvpár (EN→DE, DE→EN, ES→DE, EN→FR, ZH→EN)

Néhány kísérleti eredmény

Toral, A (2019). Post-editese: an Exacerbated Translationese. In *Proceedings of Machine Translation Summit XVII*, pp. 273–281

- Az ember által javított gépi fordítást vizsgálja (Post-editing, PE)
- Három adathalmaz (újsághírek, filmfeliratok), öt nyelvpár (EN→DE, DE→EN, ES→DE, EN→FR, ZH→EN)
- **A PE szókincse kevésbé diverz és kevésbé "sűrű", mint az emberi fordításé** (de jobb, mint a sima MT)

Angol-magyar kísérletek

Angol-magyar kísérletek

Recski, G., & Kádár, F. (2023). Language complexity in human and machine translation: a preliminary study. In: *Proc. of the International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*, pp. 268–281.

Angol-magyar kísérletek

Forrás		Átl. mondathossz	Szókincs mérete
1984	Orwell: *1984*	23.96	1428
TED	2 TED előadás leirata	15.52	990
FGM	A *Few Good Men* c. film leirata	9.62	1106
DC567	Az EU Bizottság egy állásfoglalása	22.29	1059

Angol-magyar kísérletek

Angol-magyar kísérletek

- Mindegyik forrásminta 4-6000 szóból áll

Angol-magyar kísérletek

- Mindegyik forrásminta 4-6000 szóból áll
- Minden adat online, nyilvános forrásokból származik (!)

Angol-magyar kísérletek

- Mindegyik forrásminta 4-6000 szóból áll
- Minden adat online, nyilvános forrásokból származik (!)
- A gépi fordítások a DeepL API használatával készültek

Measure of Textual Lexical Diversity (MTLD)

Measure of Textual Lexical Diversity (MTLD)

Type-token ratio:

Measure of Textual Lexical Diversity (MTLD)

Type-token ratio:

$$TTR = \frac{\log \text{különböző szavak száma}}{\log \text{összes szavak száma}}$$

Measure of Textual Lexical Diversity (MTLD)

Type-token ratio:

$$TTR = \frac{\log \text{különböző szavak száma}}{\log \text{összes szavak száma}}$$

Ez így nagyon érzékeny a hosszra.

Measure of Textual Lexical Diversity (MTLD)

Type-token ratio:

$$TTR = \frac{\log \text{különböző szavak száma}}{\log \text{összes szavak száma}}$$

Ez így nagyon érzékeny a hosszra.

McCarthy, P.M., Jarvis, S. (2010): MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* **42**(2), pp. 381–392

Measure of Textual Lexical Diversity (MTLD)

Type-token ratio:

$$TTR = \frac{\log \text{különböző szavak száma}}{\log \text{összes szavak száma}}$$

Ez így nagyon érzékeny a hosszra.

McCarthy, P.M., Jarvis, S. (2010): MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* **42**(2), pp. 381–392

MTLD (Measure of Textual Lexical Diversity)

Measure of Textual Lexical Diversity (MTLD)

Type-token ratio:

$$TTR = \frac{\log \text{különböző szavak száma}}{\log \text{összes szavak száma}}$$

Ez így nagyon érzékeny a hosszra.

McCarthy, P.M., Jarvis, S. (2010): MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* **42**(2), pp. 381–392

MTLD (Measure of Textual Lexical Diversity)

Átlagosan hány szavanként megy a TTR egy fix küszöb (0.72) alá.

Eredmények

		w/s	M_w	M_l
1984	HT	20.08	105.6	72.6
	MT	19.75	82.1	58.6
TED3	HT	13.71	67.1	46.5
	MT	14.02	55.2	38.2
FGM	HT	6.55	54.4	40.0
	MT	5.91	38.9	30.9
DC567	HT	22.48	109.7	71.4
	MT	22.13	93.9	59.2

Mondathossz és komplex szavak