

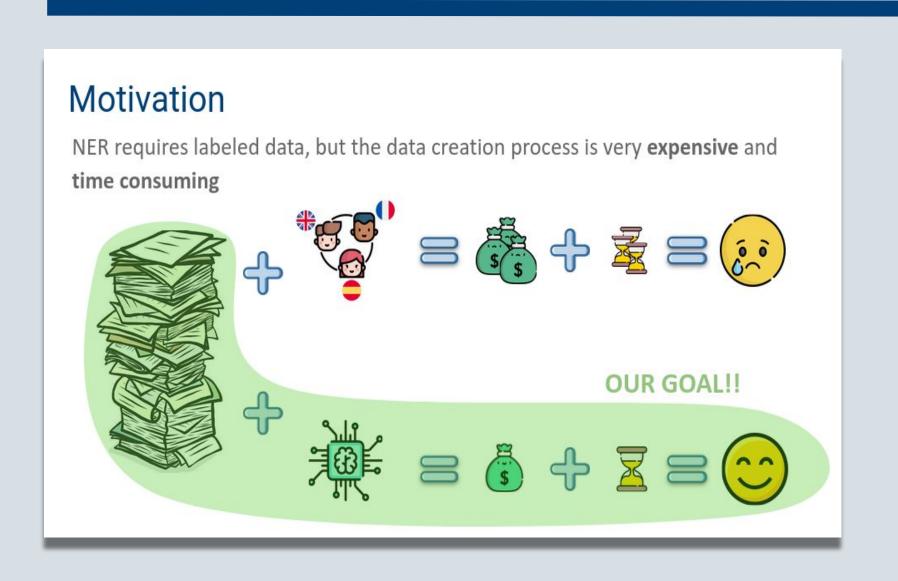


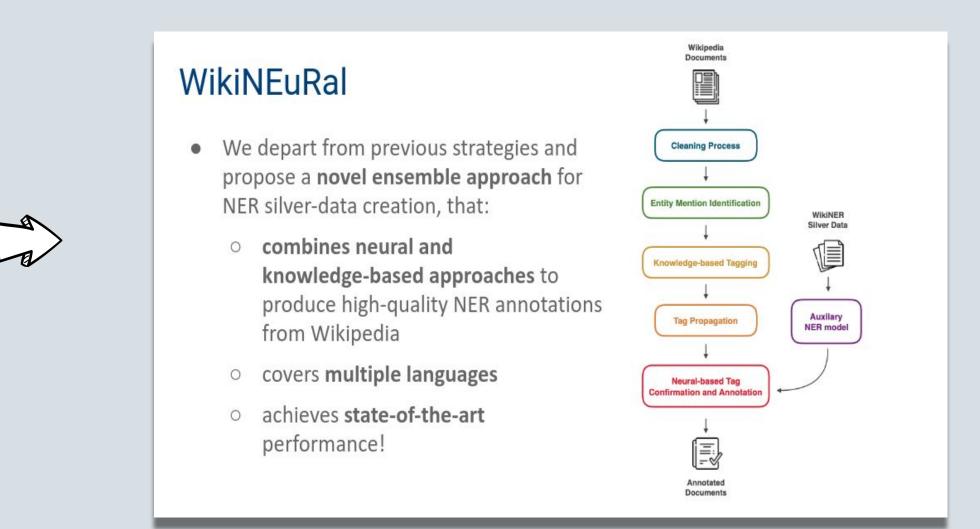
On the Intra- and Inter-linguistic Challenges of Multilingual Silver-Data Creation and Disambiguation Biases

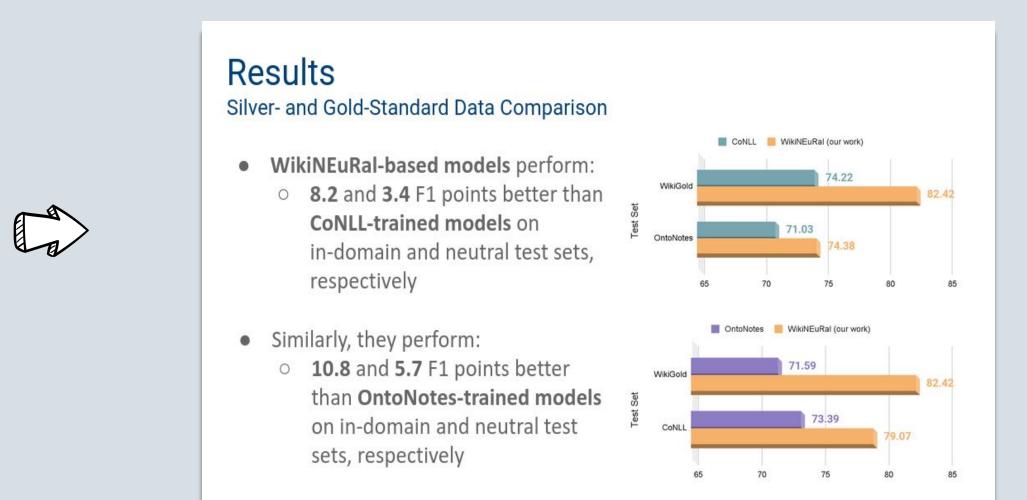
Edoardo Barba¹, Niccolò Campolungo¹, Simone Tedeschi^{1,2} and Roberto Navigli¹

¹Sapienza NLP Group, Sapienza University of Rome ²Babelscape

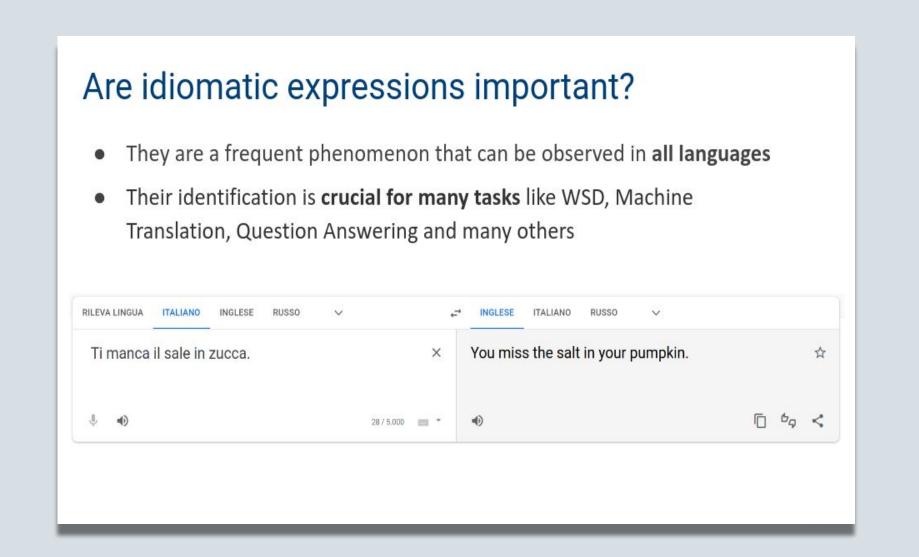
Multilingual Silver-Data Creation for Named Entity Recognition [1]

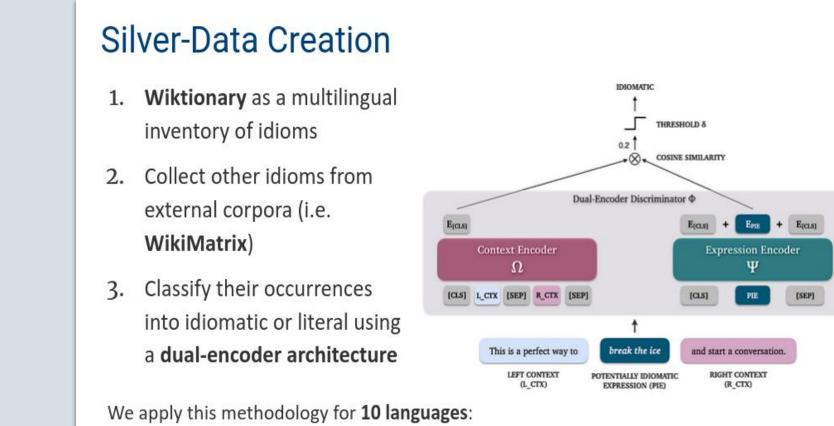






...and Idiomatic Expression Identification [2]



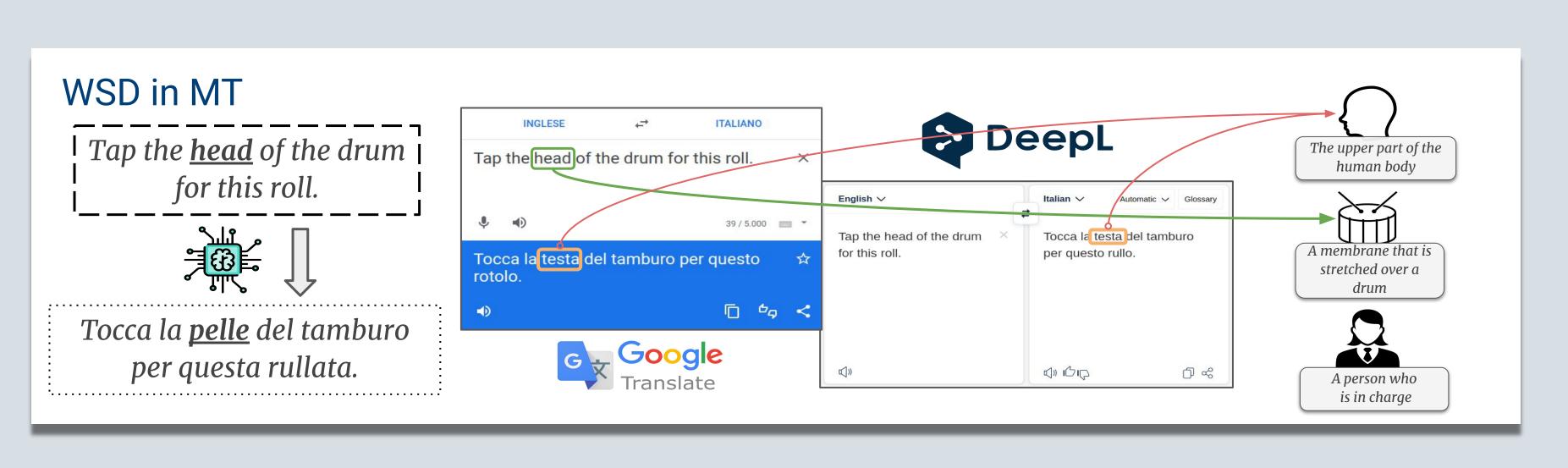


[1] Tedeschi et al. (2021) "WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER."

de, en, es, fr, it, ja, nl, pl, pt, zh

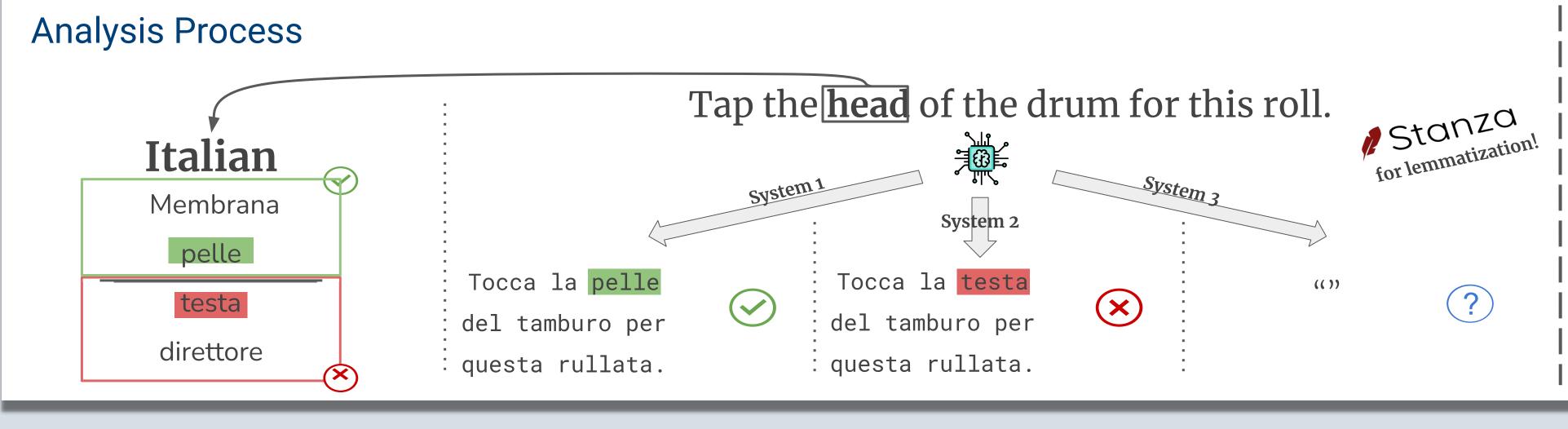
[2] Tedeschi et al. (2022) "ID10M: Idiom Identification in 10 Languages."

Disambiguation Biases in Machine Translation [3]





From WordNet and Wiktionary we selected max one sentence per sense per source Each selected sentence contains a target word with polysemy degree > 2 and a sufficient semantic context Target words are single-token and MWEs but not idioms





[3] Campolungo et al. (2022) "DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation."

TL;DR

- Silver-data creation is a powerful, fast and cheap tool that can be used to tackle both inter- and intra-linguistic challenges in NLP by producing training data for:
- Low-resource languages
- A variety of tasks, including those involving figurative language
- Lexical-semantic disambiguation biases strongly affect NLP systems
- Analyses on the DiBiMT benchmark show that MT models are still far from correctly handling infrequent senses

• Relevant WGs: 1, 3 and 4

Reach out!

@edoardo_barba @Valahaar @SimoneTedeschi_ @RNavigli













