UniDive 1st general meeting

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

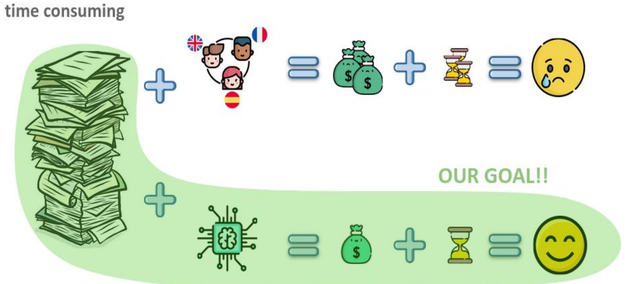
Edoardo Barba[1], Niccolò Campolungo[1], Simone Tedeschi[1,2] and Roberto Navigli[1]

[1]Sapienza NLP Group, Sapienza University of Rome [2]Babelscape

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

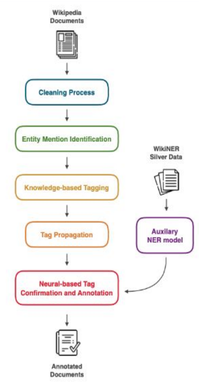
Motivation

NER requires labeled data, but the data creation process is very **expensive** and **time consuming**



WikiNEuRal

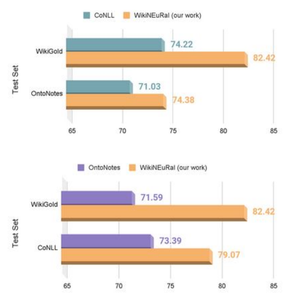
- We depart from previous strategies and propose a **novel ensemble approach** for NER silver-data creation, that:
  - **combines neural and knowledge-based approaches** to produce high-quality NER annotations from Wikipedia
  - covers **multiple languages**
  - achieves **state-of-the-art** performance!



Results

Silver- and Gold-Standard Data Comparison

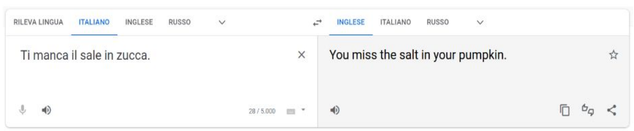
- **WikiNEuRal-based models** perform:
  - **8.2** and **3.4** F1 points better than **CoNLL-trained models** on in-domain and neutral test sets, respectively
- Similarly, they perform:
  - **10.8** and **5.7** F1 points better than **OntoNotes-trained models** on in-domain and neutral test sets, respectively



...and Idiomatic Expression Identification [2]

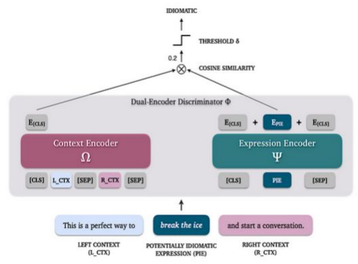
Are idiomatic expressions important?

- They are a frequent phenomenon that can be observed in **all languages**
- Their identification is **crucial for many tasks** like WSD, Machine Translation, Question Answering and many others



Silver-Data Creation

1. **Wiktionary** as a multilingual inventory of idioms
2. Collect other idioms from external corpora (i.e. **WikiMatrix**)
3. Classify their occurrences into idiomatic or literal using a **dual-encoder architecture**



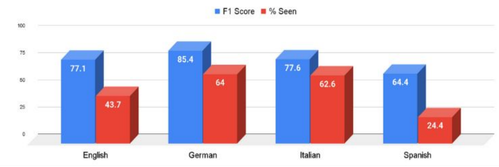
We apply this methodology for **10 languages**: de, en, es, fr, it, ja, nl, pl, pt, zh

Task Reformulation and Results

- **Task Reformulation:** in the previous formulation, the expression was pre-identified

Tokens	*After*	*some*	*reflection*	*,*	*he*	*decided*	*to*	*bite*	*the*	*bullet*	*.*
Tags	O	O	O	O	O	O	O	B-IDIOM	I-IDIOM	I-IDIOM	O

- **Results:** 76 F1 points, % Seen entities: 48.7% (on average)

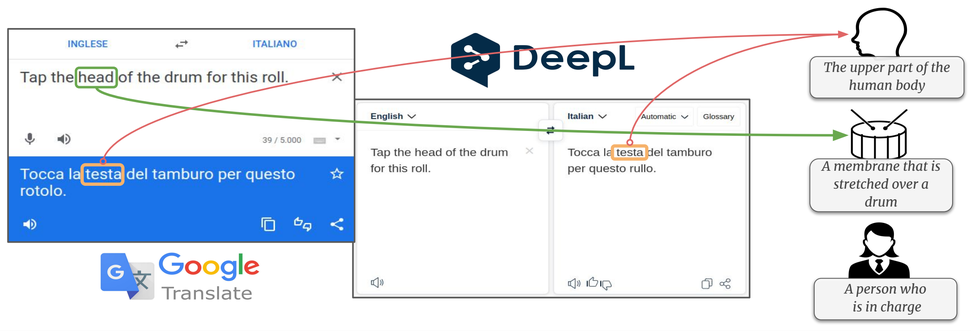


Disambiguation Biases in Machine Translation [3]

WSD in MT

*Tap the **head** of the drum for this roll.*

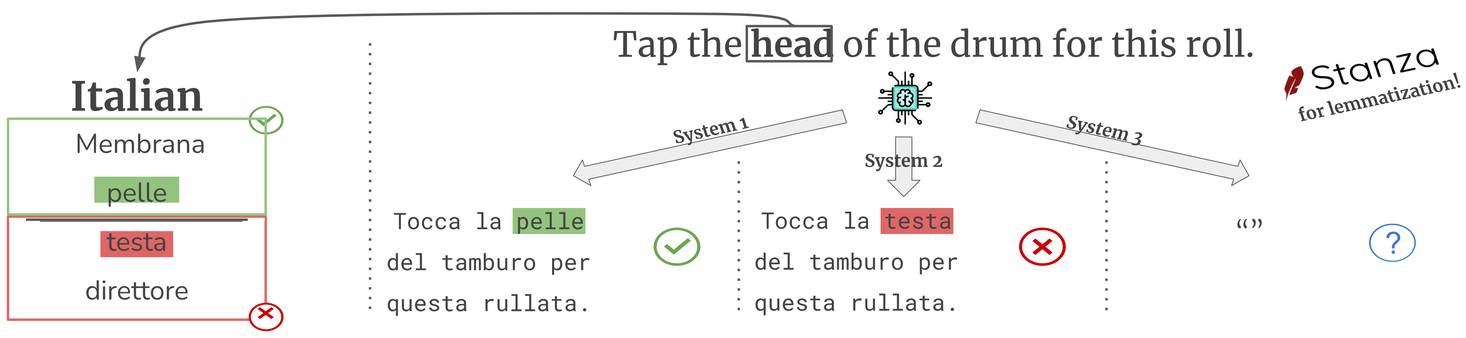
*Tocca la **pelle** del tamburo per questa rullata.*



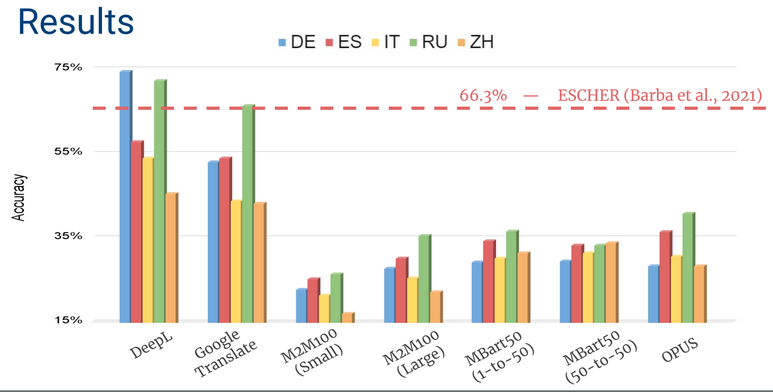
Creating DiBiMT

- 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**

Analysis Process



Results



[1] Tedeschi et al. (2021) *"WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER."* [2] Tedeschi et al. (2022) *"ID10M: Idiom Identification in 10 Languages."*

[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

Project No. 731015 (elexis european lexicographic infrastructure)
PerLIR Project (PRIN MIUR)
MOUSSE ERC Consolidator Grant No. 726487
Cost Action CA21167 (UniDive)