Unsupervised MT for under-resourced languages

Jordi Armengol-Estapé

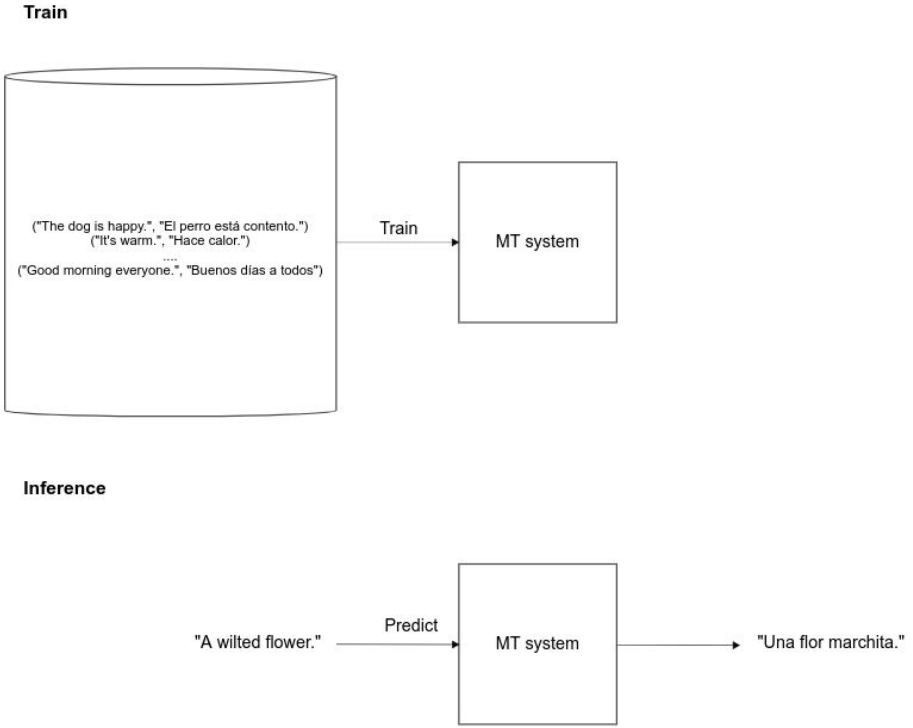
@jordiae 

jordi.armengol.estape@ed.ac.uk

25/06 - SIGUL - LREC 2022

Machine Translation

Machine Translation (MT) systems are trained with *lots* of sentence pairs:



Unsupervised MT is a perfect match for under-resourced languages

Machine Translation (MT) systems are trained with *lots* of sentence pairs.

However, this data is *not* generally available for under-resourced languages.

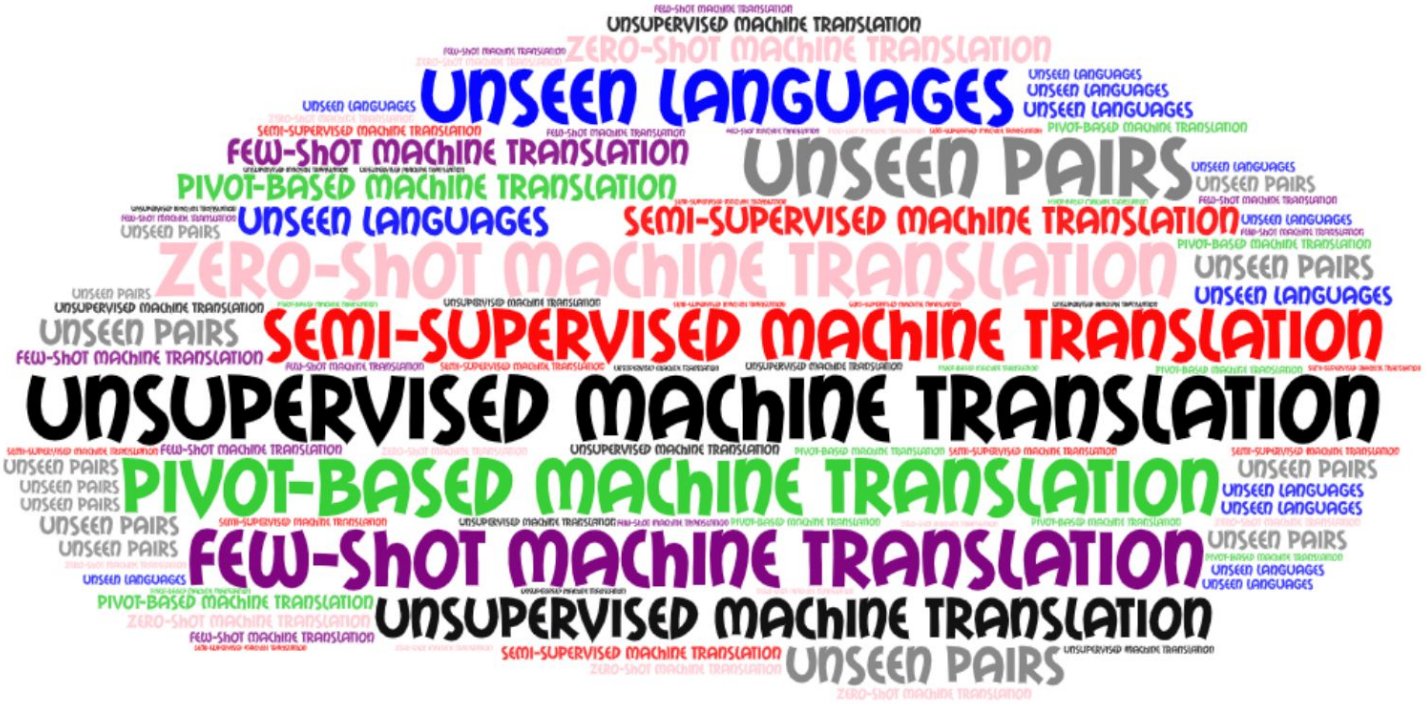
Thus, *unsupervised* MT is a perfect match for under-resourced languages!

Easy, right?

But wait…

1. What *exactly* is unsupervised machine translation?
2. Which languages are "*under resourced*" in machine translation?
3. Why should we care about unsupervised translation for under-resourced languages (*should we?*)?

What is unsupervised MT?



What is unsupervised MT?

Unsupervised: translate Source into Target having no (0) labeled examples.

What is unsupervised MT?

Unsupervised: translate Source into Target having no (0) labeled examples.

But **not** having labeled examples *of what?*

A. Of pairs Source-Target, but we have monolingual Source and Target?
   [Assuming we can identify them! *Rethinking the Truly Unsupervised Image-to-Image Translation (Baek et al., 2020)*]
B. Of pairs Source-Target, but we have Source-Foo and Bar-Target?
C. Of pairs Source-Target, but we have monolingual Source and Bar-Target?
D. Of Source, but we have Bar-Target, and Bar is similar to Source?
E. Of Target?

What is unsupervised MT?

Unsupervised: translate Source into Target having no (0) labeled examples.

But not having labeled examples *of what?*

A. Of pairs Source-Target, but we have monolingual Source and Target? **Artetxe, Lample (XLM)**
(Assuming we can identify them! *Rethinking the Truly Unsupervised Image-to-Image Translation (Baek et al., 2020)*

B. Of pairs Source-Target, but we have Source-Foo and Bar-Target?
C. Of pairs Source-Target, but we have monolingual Source and Bar-Target?
D. Of Source, but we have Bar-Target, and Bar is similar to Source?

Multilingual models

~~E. Of Target?~~ **Impossible**

All these are legit instances of unsupervised MT!

Today, we assume bilingual settings (scenario A).

1. ~~What *exactly* is unsupervised translation?~~
2. Which languages are “*under resourced*” in machine translation?
3. Why should we care about unsupervised translation for under-resourced languages (*should we?*)?

What do we mean by under-resourced languages?

NLP in general:

- Unclear where to put the bar (10k sentences? 100k? 1M?)

MT-specific:

- Data *relative to the difficulty* (language similarity).
- Low-resource *pairs:*
  - e.g. large English-Spanish and English-Russian datasets, but not so large Russian-Spanish?
- Domains!
  - e.g. not so many resources for biomedical Spanish!

1. ~~What *exactly* is unsupervised translation?~~
2. ~~Which languages are "*under resourced*" in machine translation?~~
3. Why should we care about unsupervised translation for under-resourced languages (*should we?*)?

Why unsupervised translation?

We got it, it’s cool for under-resourced languages…

But seminal unsupervised MT works use general-domain English, German as benchmarks!

What about real-world scenarios?

MT4All: Unsupervised Machine Translation in Real-world Scenarios (de Gibert Bonet, 2022)

Why unsupervised translation?

A Call for More Rigor in Unsupervised Cross-lingual Learning (Artetxe et al., 2020):

“We argue that a scenario without any parallel data and abundant monolingual data is unrealistic in practice.”

Current trends:

- Purely unsupervised is not compulsory! (I’m pretty sure there are supervised translation examples in GPT-3’s corpus)
- Massively multilingual models
- Transfer learning from *big* models

Takeaways

Scientifically, unsupervised machine translation *does* matter!

(It drives innovation relevant to under-resourced languages.)

In practice, sticking to pure “unsupervision” might be:

- a detrimental, self-imposed challenge if semi-supervised methods are possible.
- a hopeless challenge if monolingual data is scarce.