Shared-Task Driven RBMT development

apertium-fin-eng experiments at WMT 2015–2019

https://github.com/apertium/apertium-fin-eng

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

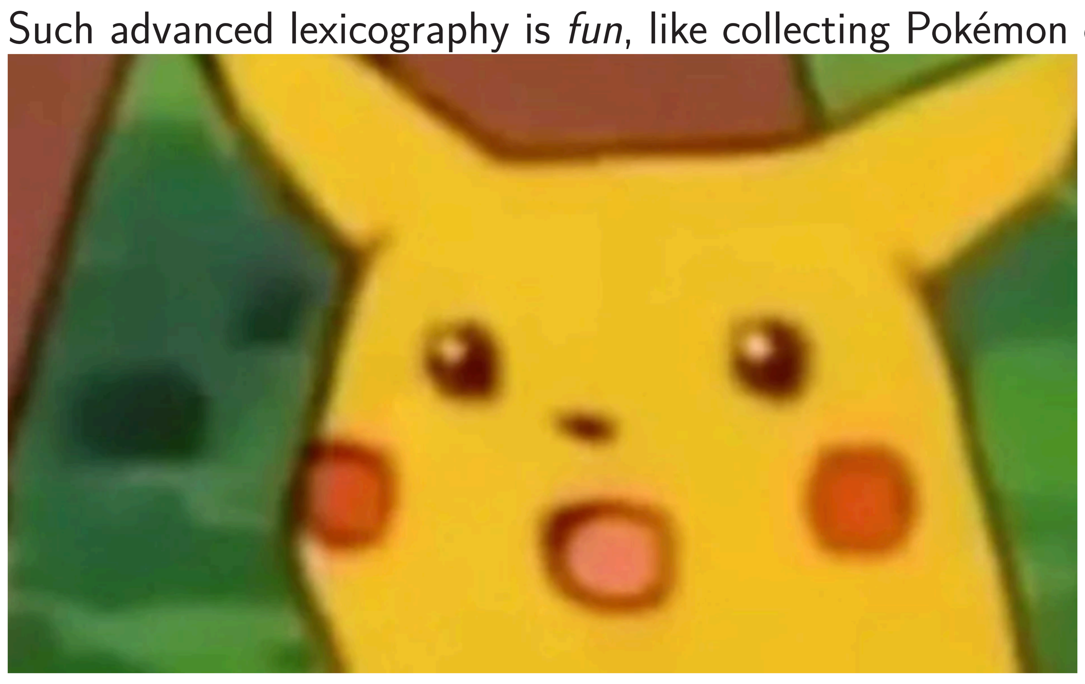
hzsk

What is this?

- Rule-based, shallow transfer, machine translation system for Finnish—English using Apertium
- A dictionary and rule development *workflow* for RBMT
- Probably not the best RBMT for fin—eng

Why? Why not?

- Advantages of rule-based systems:
  - Such advanced lexicography is *fun*, like collecting Pokémon or stamps!



  - Working on the rules is also *fun*, like using your brain to understand linguistics and all!
  - Does not destroy our planet at an alarming pace! (c.f. [SGM19] few days earlier in this ACL 2019 ;-)

  - Does not need to leak your data through internet and to big companies!
  - Predictable errors, easy fixes!
- Some disadvantages of Shallow RBMT here:
  - BLEU score always low
  - Shallow transfer RBMT not particularly suitable for Finnish—English; needs more depth
  - Finnish—English is a very well resourced language pair indeed, so simply SMT and NMT makes a lot of sense
- Therefore:
  - I mainly use work on fin—eng as an additional data point on measuring workflow efficiency
  - come to Dublin for MTsummit / LoResMT workshop to see another data point with interestinger languages [Pir19] (no English!)

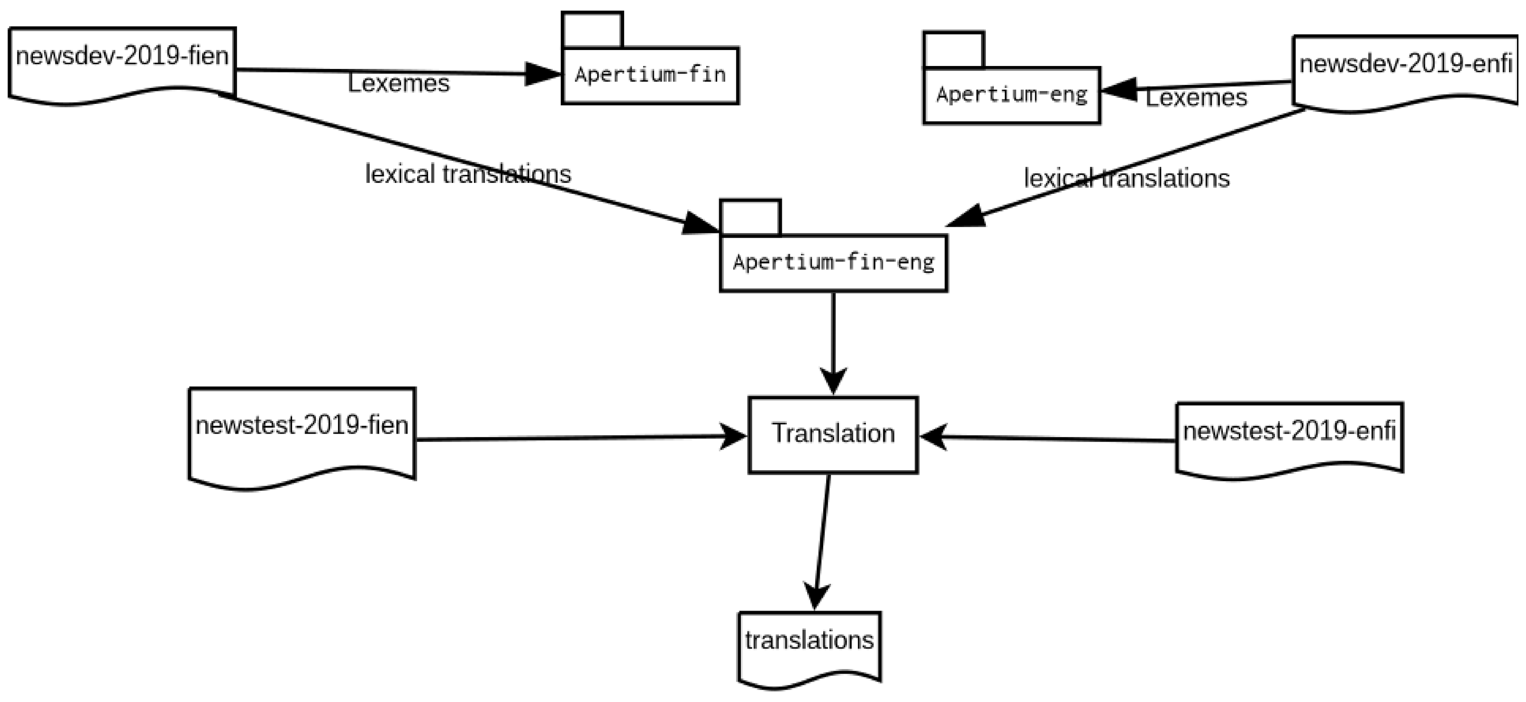
How do you do it?

- It's a workflow, based on shared tasks, so *Shared Task Driven* development
- Do this:
  1. Collect all lexemes unknown to source language dictionary, and add them with necessary morpholexical information
  2. Collect all lexemes unknown to bilingual translation dictionary, and add their translations
  3. Collect all lexemes unknown to the target language dictionary, and add them to the dictionary with necessary morpholexical information
- repeat until all words from shared task are in all dictionaries
- Can be semi-automated to great extent
- Grammar rules, on the other hand, are mostly manual, expert labour still

References

- Tommi A Pirinen.
  Workflows for kickstarting rbmt in virtually no-resource situation.
  In *Proceedings of The 2nd Workshop on Technologies for MT of Low Resource Languages (LoResMT 2019)*, Dublin, Ireland, 2019.
  to appear.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum.
  Energy and policy considerations for deep learning in NLP.
  In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics.

Figure



Results

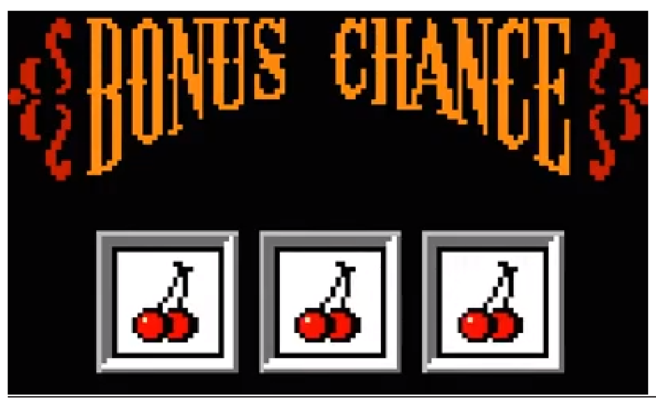
Error	count
OOVs in Finnish	763
OOVs in English	943
OOVs in Fin Eng	2696

Table: Classification of mainly lexical errors in apertium-fin-eng submissions for 2019

Corpus		BLEU-cased
apertium-eng-fin	2015	2.9
	2017	3.5
	2019	4.3
apertium-fin-eng	2015	6.9
	2017	6.3
	2019	7.6

Table: Progress of apertium-fin-eng over the years using only the WMT shared task driven development method.

Some cherry-picked examples for fun and entertainment



Source	Aika nopeasti saatiin hommat sovittua, Kouki sanoi
Apertium-fin-eng	Kinda swiftly let jobs agreed, Kouki said.
an NMT at WMT2019	Pretty quickly we got the gays agreed, Kouki said.
Reference	We reached a pretty quick agreement, Kouki said.

Source	Natural disasters make logistics even more complicated.
Apertium-eng-fin	Luontevat tuhot malli- logistiikka vielä enemmän sekava.
(my re-translation)	Natural destruction model- logistics even more confusing.
an NMT at WMT 2019	Luonnonkatastrofit tekevät saasteista entistä monimutkaisempia.
(my re-translation)	Natural disasters make pollution even more complicated
Reference	Luonnonkatastrofit tekevät logistiikasta vieläkin monimutkaisempaa.

Acknowledgments

The author was employed by CLARIN-D during WMT 2019. The free/libre open source RBMT systems have been made possible by contributors of omorfi, apertium-fin, apertium-eng and apertium-fin-eng packages.

Apertium