Parallel Corpora in (Machine) Translation: goals, issues and methodologies

Time in Translation Workshop

University of Utrecht, 23rd June 2017

Antonio Toral
a.toral.ruiz@rug.nl

Content

1. Introduction: Why, Which, Where, How
2. Massive Acquisition Made Easy
3. Cleaning the Messy
4. Creative Translators and Parallel Corpora

Context, about me

- Named Entities & WSD, till 2009
  - PhD at Univ. Alacant & CNR Pisa
  - Topic: Named Entity Acquisition from Wikipedia
- MT since 2010
  - Researcher in MT at DCU
  - Assistant Prof at Univ. Groningen (since October 2016)
  - Topics
    - Corpora acquisition
    - Domain adaptation
    - Diagnostic evaluation
    - Literary text

1 Introduction:
Why, Which, Where, How

Why use parallel corpora

- Machine Translation (MT)
  - Statistical approaches (phrase-based, neural): key component
  - Rule-based approaches: automatic acquisition of rules and dictionaries
    - Less manual effort required
    - Resulting rules and dictionaries reflect language use
- Computer-assisted translation
  - Translation memories, fuzzy matches
- Corpus-based Linguistic Research of Translations

Which Parallel Corpora

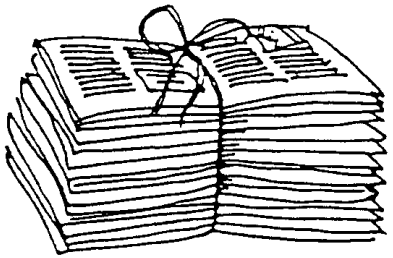
- Wish list
  - Clean
  - Appropriate domain, genre and style
  - Big size
  - ...

Where to get parallel corpora

- Already existing
  - free (Europarl, OPUS) versus
  - for a fee (ELDA, LDC, TAUS, etc)
- Web crawling
  - ad-hoc (e.g. SETimes, TLAXCALA) versus
  - generic (e.g. Bitextor, ILSP Focused Crawler)
- Generate manually
  - Professional translation (~0.05 cent/word) versus
  - crowdsourcing (very cheap but: quality, ethics...)
  - Trade-off cost-benefit: careful selection of data to be translated
    - Variety: avoid redundance in sentences translated
    - Performance: predict which data, if it were translated, would improve (MT) the most

How to format it

- Tokenised
  - and case-normalised (truecased, lowercased)
- Sentence aligned
  - E.g. Hunalign (2005)
- Word aligned?
  - E.g. GIZA++, fast_align
- ...



2 Massive acquisition... made easy

Massive Acquisition

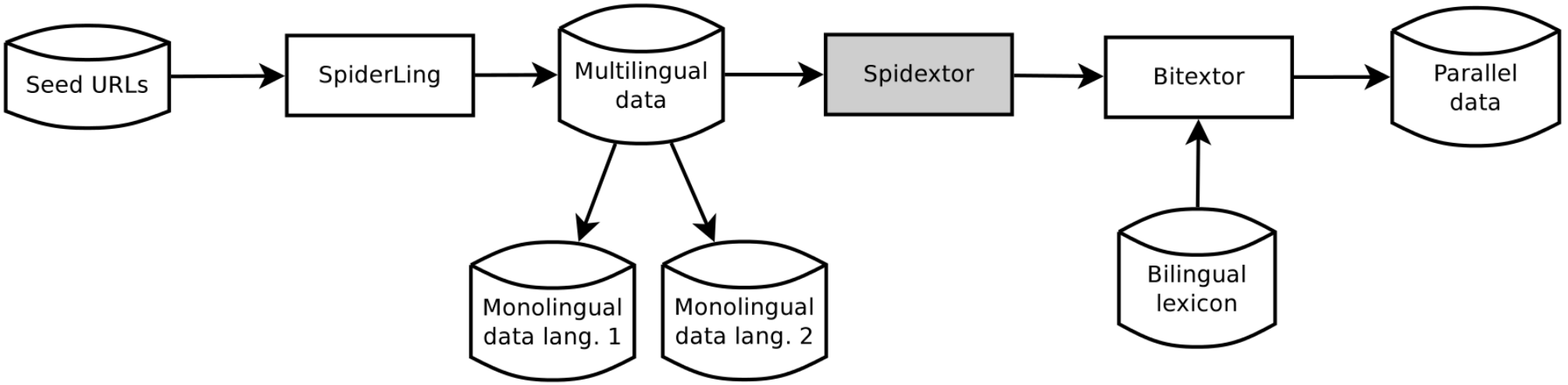
- Current parallel crawlers
  - Input: URL of a web domain with parallel data
  - Output: parallel documents identified from that web domain
- Problem
  - One needs to compile the list of promising web domains manually!
- Solution
  - Automatically identify promising web domains

Massive Acquisition

- Spidextor [1]
  - Crawling of parallel (and monolingual) corpora from top level domains
  - Example
    - Inputs
      - TLD: .nl
      - Languages of interest: en and nl
    - Output: corpora crawled
      - Monolingual: en, nl
      - Parallel: en-nl

[1] Ljubešić et al. Producing monolingual web corpora and bitext at the same time -- SpiderLing and bitextor's love affair. LREC 2016.

Massive Acquisition



Results

- 4 TLDs: .fi, .hr, .sl, .sr
- Crawled for 3 to 7 days

corpus	web domains	segments	words
fienWaC	10,664	2,866,574	77,048,083
hrenWaC	5,624	1,554,912	55,083,246
slenWaC	3,529	718,315	27,924,210
srenWaC	2,546	534,682	23,139,804

Evaluation

- Intrinsic (% non parallel segments)
  - 16% (en-hr) to 32% (en-sl) noise
- Extrinsic

direction	system	BLEU	TER
en→hr	Google	0.2673	0.5946
	Bing	0.2281	0.6263
	Yandex	0.2030	0.6801
	hrenWaC	0.2457	0.6198
	all	0.2445	0.6147
hr→en	Google	0.4099	0.4635
	Bing	0.3658	0.5199
	Yandex	0.3463	0.5311
	hrenWaC	0.3499	0.5090
	all	0.3721	0.4878

Issues (I): Quality

- Is web crawled parallel data not low-quality?
  - **Noise**: documents in languages A and B that are not translations of each other
    - Sentence alignment confidence scores (sentence and document level)
      - Where to set the threshold: precision vs recall trade-off
  - Some data may not be human but **machine translation**!
    - Tell apart human from machine translated documents with a binary classifier

Issues (II): Domain

- Web crawled corpora from TLD (W) represents the “general” domain...
- ... but I am interested in a particular domain (D)

- Find the subset of W that belongs (or is very related) to D
  - Rank sentences (or documents) in W according to their similarity to a small corpus in the domain D

3 Cleaning the messy

Motivation

- Many publicly available parallel corpora are potentialy useful
- But... they are too noisy
  - Missalignments
  - Encoding errors
  - etc

Case Study: OpenSubtitles

- Available for many language pairs
- Big size
- But very messy

- How can we unlock its potential?

Procedure

- Automatic cleaning
  - Fixing (sparsity)
  - Removing sentences (noise)

[2] Forcada et al. D4.1b MT systems for the second development cycle. Abu-MaTran deliverable. 2014.

Procedure

- Automatic cleaning
  - Fixing (sparsity)
    - Converting Cyrillic characters to their Latin counterparts
    - Converting encoding to UTF-8
    - Spelling errors
    - Inconsistent punctuation marks, numbers and spacing
  - Removing sentences (noise)
    - Without alphabetical characters
    - Too different in length
    - Not in the right language

[2] Forcada et al. D4.1b MT systems for the second development cycle. Abu-MaTran deliverable. 2014.

Results

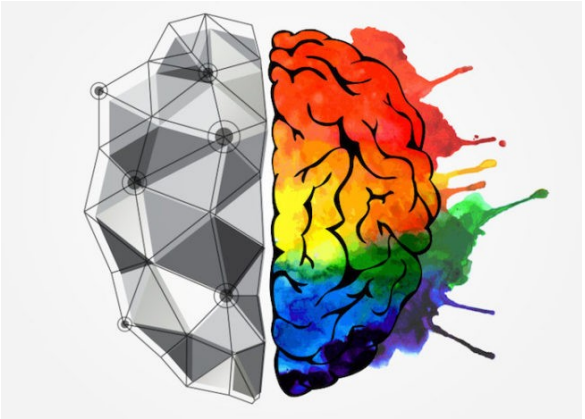
- Data
  - Corpora: OpenSubtitles en-hr
  - Input: 30M sentence pairs
  - Output: 17M

- Extrinsic Evaluation
  - Train MT system with OpenSubs as is vs cleaned
  - Test set: news domain

Results

- MT results (BLEU)

	EN-HR	HR-EN
OpenSubs as is	0.09	0.22
OpenSubs cleaned	0.22	0.31
Relative improvement	145%	37%



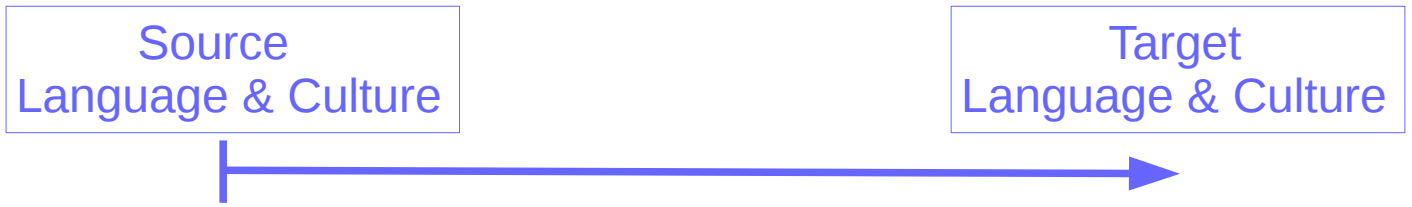
4 Creative translators and parallel corpora

Translation Options

- There are many possible translations
- What makes a good translation?
- Different views in Translation Studies, e.g:
  - Domesticating. Bring the original to the target audience
  - Foreignising. Bring the target audience to the source text

Translation Options

- There are many possible translations
- What makes a good translation?
- Different views in Translation Studies, e.g:
  - Domesticating. Bring the original to the target audience
  - Foreignising. Bring the target audience to the source text



Translation Options

Lui parti, j’ai retrouvé le calme.
J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.
Des bruits de campagne montaient jusqu’à moi.
Des odeurs de nuit, de terre et de sel rafraîchissaient mes tempes.
La merveilleuse paix de cet été endormi entrait en moi comme une marée.
A ce moment, et à la limite de la nuit, des sirènes ont hurlé.
Elles annonçaient des départs pour un monde qui maintenant m’était à jamais indifférent.
Pour la première fois depuis bien longtemps, j’ai pensé à maman.

Jones and Irvine (2013)
Toral and Way (2015)

Translation Options

French
J’étais épuisé et je me suis jeté sur ma couchette. Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.
English – translation A
But all this excitement had exhausted me and I dropped heavily on to my sleeping plank. I must have had a longish sleep, for, when I woke, the stars were shining down on my face.
English – translation B
I was exhausted and threw myself on my bunk. I must have fallen asleep, because I woke up with the stars in my face.

Is any of the translations better/easier for MT and/or corpus studies?

Translation Options

French

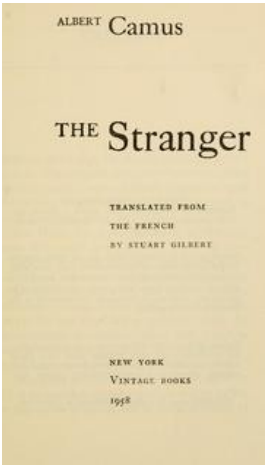
J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.

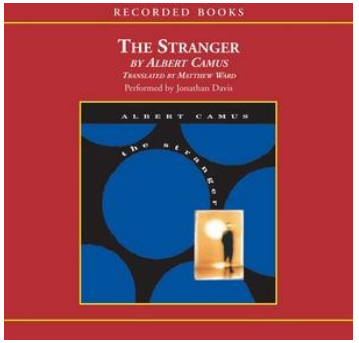
English – Gilbert (1946)

But all this excitement had exhausted me and I dropped heavily on to my sleeping plank.
I must have had a longish sleep, for, when I woke, the stars were shining down on my face.

English – Ward (1989)

I was exhausted and threw myself on my bunk.
I must have fallen asleep, because I woke up with the stars in my face.





Translation Options

French

J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.

English – Gilbert (1946)

But all this excitement had exhausted me and I dropped heavily on to my sleeping plank.
I must have had a longish sleep, for, when I woke, the stars were shining down on my face.

English – Ward (1989)

I was exhausted and threw myself on my bunk.
I must have fallen asleep, because I woke up with the stars in my face.

Domesticating
Transcreation
Free translation

Foreignising
Literal translation

Translation Options

French

J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.

English – Gilbert (1946)

But all this excitement had exhausted me and I dropped heavily on to my sleeping plank.
I must have had a longish sleep, for, when I woke, the stars were shining down on my face.

BLEU 0.11
TER 0.80

English – Ward (1989)

I was exhausted and threw myself on my bunk.
I must have fallen asleep, because I woke up with the stars in my face.

BLEU 0.28
TER 0.56

[3] Toral and Way. Machine-Assisted Translation of Literary Text: A Case Study. Translation Spaces. 2015.

Consequences

- Strategies such as domestication and transcreation lead to translations that can differ significantly from the source text
- This can be challenging for corpus-based research Examples:
  - MT. Poor performance due to difficulty to
    - Find reliable word alignments (training)
    - Find n-gram matches (automatic evaluation)
  - Finding corresponding tense
    - If they are the same in both languages
      - can this be attributed to similarity of use in both languages, or to the translation being foreignising?
    - If they are different
      - can this be attributed to a divergence of preferred tense between both languages, or to the stylistic choice of the translator?

5 Conclusions

1. Web crawling

- Easy to use automatic procedure to acquire massive amounts of parallel data
- Can be paired with post-processing to
  - Remove noise: missalignments, MT
  - Find data for a specific domain of interest

2. Cleaning

- It is possible to clean very messy corpora...
- ...at the expense of removing a lot of data
- Opensubs use case
  - Removed: 43%
  - MT scores increase
    - HR -> EN 37%
    - EN -> HR 145%

3. Translation strategies

- Be aware that human translations can differ substantially based on strategies, theories, etc.
- One important variable to take into account when using parallel corpora
- Should we do anything about it?
  - e.g. identify the translation strategy used for each parallel corpus

Acknowledgements

- Massive Acquisition
- Cleaning corpora

Acknowledgements

- Massive Acquisition
- Cleaning corpora

АБУМАТРАН
ABUMATRAN

- Creative translators

PiPeNovel

The End! Questions?

Parallel Corpora in (Machine) Translation: goals, issues and methodologies

23rd June 2017

Antonio Toral
a.toral.ruiz@rug.nl