NMT for (really) low-resource languages

Robert Östling
robert@ling.su.se

2018-11-20

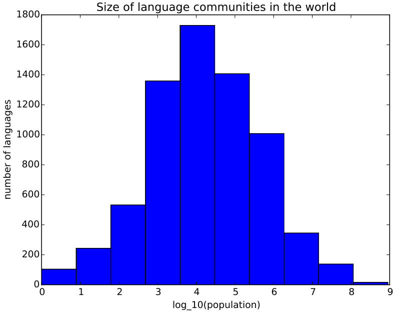
Stockholm University

About myself

- At the Department of **Linguistics**
- Practical MT is fun, and so is (impractical?) linguistics
- Particular interest in highly multilingual NLP+linguistics

Languages of the world

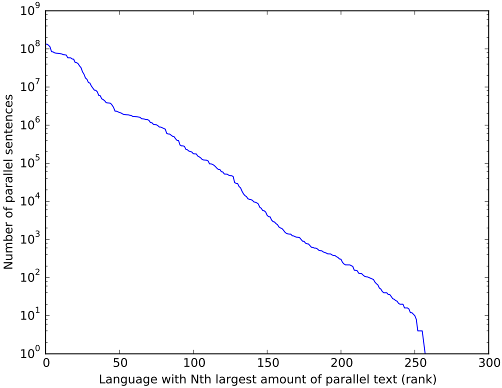
- Nearly all languages are small
- If we are **lucky**, there might be some subset of:
  - a grammatical description
  - a lexicon (often just a 100-word Swadesh list)
  - bits of text online (say, a Facebook group)
  - a small corpus by a linguistic fieldworker
  - a few translated texts



MT resources

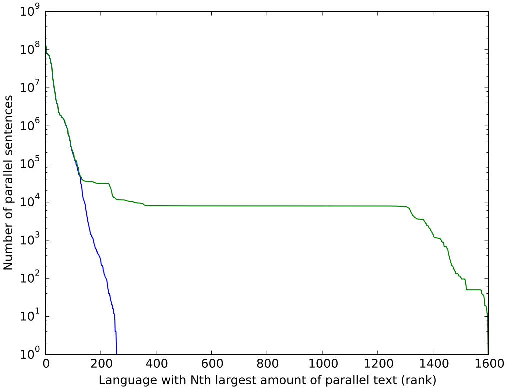
What do the standard MT resources look like?

OPUS



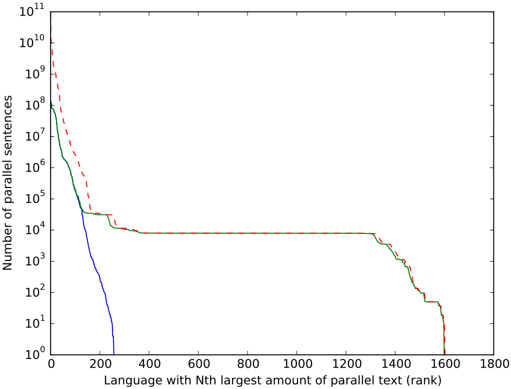
Coverage: **one language in 30**

OPUS + Marburg



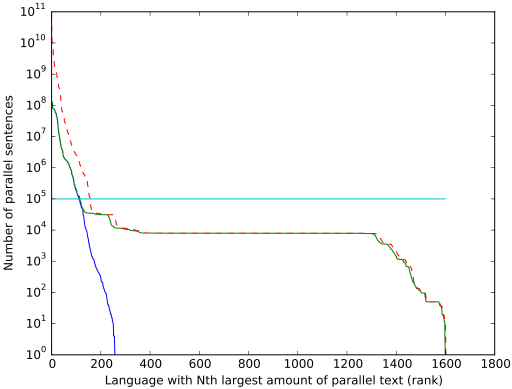
Coverage: **one language in 5**

OPUS + Marburg + CommonCrawl



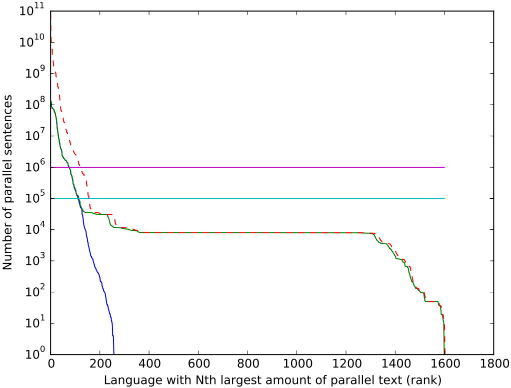
For most languages **monolingual** ≈ **parallel**

Data limits of Machine Translation: parallel



MT limit ≈ 100 000 **parallel sentences**

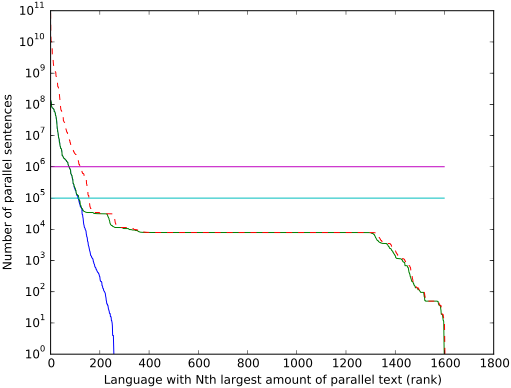
Data limits of Machine Translation: monolingual



MT limit $\approx$ 1 000 000 **monolingual sentences(?)**

Some initial conclusions

- Supervised MT supports 100–200 languages
- Unsupervised MT supports 100–200 languages
- Parallel text mining supports these 100-200 languages
- Rule-based MT supports all languages, but would cost a bit



So what can we do?

- Obtain more data
- Improvements in "data efficiency" (BLEU per megasentence)
- Marginal gain: one order of magnitude $\approx$ 40 more languages (assuming quality requirements are constant)
- ...**until** you reach 10^4 when you get a bonus of 1000 languages!
- Is this possible? (So far NMT has not been very helpful)

Upper limit

- Typical data: New Testament translation in some small language
  - Length: $\approx 10^5$ words
  - Vocabulary: ≈ 4000 lemmas, specific domain
- What could a (computer-aided) human linguist do?
  - Learn most of the grammar
  - Learn the basic lexicon
  - Make educated guesses for unseen words based on word formation patterns, cognates, loan words, typical patterns of polysemy
  - Use world knowledge to guess the meaning of unclear texts
  - Express this hypothesis in some other language, i.e. translation
- Do we need non-traditional MT data sources?

Kinds of data not frequently used

- From (grammatical) typology: “house tree destroy”
- From lexical typology: “my house was destroyed by a tree”
  Hint: polysemy patterns involving ‘tree’
- From historical linguistics: sound changes, cognates
- From English/Big Data: world knowledge (hopefully)

Let's get on with the NLP

1. Multilingual word representations (with Murathan Kurfalı)
2. Language representations (with Jörg Tiedemann)

Multilingual word embeddings

- A standard building block of multilingual NLP
- We want these properties:
  1. $d(\mathrm{dog}, \mathrm{cat}) < d(\mathrm{dog}, \mathrm{apple})$
  2. $d(\mathrm{dog}, \mathrm{Hund}) < d(\mathrm{dog}, \mathrm{Katze})$
  3. $d(\mathrm{dog}, \mathrm{Hund}) < d(\mathrm{dog}, \mathrm{cat})$
- Whether (3) is desirable depends on the application
- Most methods are designed for the top 100–200 languages
- These include:
  - learning from multilingual context
  - aligning monolingual embeddings
- Beyond the top 100–200, we can do projection through word alignments

Method

- First, note that the data is highly multi-parallel!
- Use a few high-resource languages (27) to:
  1. Learn high-quality monolingual embeddings (fastText)
  2. Align the embeddings using bilingual wordlists (Smith et al. 2017 or your favorite method)
- Word align 168 $\times$ 1 407 pairs of Bible translations
- Project high-resource embeddings to low-resource languages $v_{\text{Kamel}} = \frac{1}{N}\left(5v_{\text{camel}} + 3v_{\text{chameau}} + 2v_{\text{kamelin}} + \ldots\right)$
- Keep the 25% most coherent word types for the projection

Bitext alignment—advertisement

- $168 \times 1\,407 = 236\,376$ alignments $\leftarrow$ lots of work!
- Each alignment better be fast
  - `https://github.com/robertostling/eflomal`
  - `fast_align` compatible but faster and better
  - IBM models with Dirichlet priors, Gibbs sampling
  - Now with arbitrary user-defined priors
  - Plug in string similarity, lexicon resources, etc.
  - ...or just pretrain on large data sets

Evaluation setup

- Our approach is not ideal for translation
- Diffucult to learn e.g. Monday $\neq$ Tuesday
- But word translation by nearest-neighbor lookup is an easy way to evaluate
- Gathering word lists **consistent with Bible orthography** requires work (or noisy heuristics), so we pretend Swedish is low-resource

Let's try it — single source

	Eng to Swe		Swe to Eng	
	p@1	p@5	p@1	p@5
ind	0.137	0.344	0.173	0.335
Smith et al. (2017)	0.501	0.686	0.525	0.722

Projection from Indonesian (high-resource) to Swedish. English is only used in evaluation. Numbers using simpler filtering method.

What about multi-source?

	Eng to Swe		Swe to Eng	
	p@1	p@5	p@1	p@5
ind	0.137	0.344	0.173	0.335
ind+fin	0.231	0.462	0.223	0.394
ind+fin+hun	0.255	0.493	0.234	0.399
ind+fin+hun+tur	**0.269**	0.501	0.235	**0.400**
ind+fin+hun+tur+est	0.267	**0.504**	**0.236**	0.395
Smith et al. (2017)	0.501	0.686	0.525	0.722

What if we choose lucky languages?

	Eng to Swe		Swe to Eng	
	p@1	p@5	p@1	p@5
nob	0.275	0.493	0.344	0.521
nob+nld	0.344	0.582	0.381	0.562
nob+nld+dan	0.368	0.605	**0.386**	**0.569**
nob+nld+dan+fin	0.389	0.615	0.373	0.552
nob+nld+dan+fin+pol	**0.400**	0.623	0.372	0.556
nob+nld+dan+fin+pol+bul	0.392	**0.626**	0.363	0.546
Smith et al. (2017)	0.501	0.686	0.525	0.722

Better if we happen to have closely related high-resource languages, of course.

“Error” analysis

Source	**police** say that the **truck** driver was not drunk at the time .
Translation	**vakterna** påstå att den **vagnen** förare hade inte drucken vid den tiden .
Glossing	the-**guards** claim that the **wagon** driver had not drunken by that time .

Keeping the semantic structure of the source embedding space becomes important here.

"Error" analysis

Source	one city has no **electricity** for months .
Translation	enda stadens har inget **belysningen** för månader .
Glossing	only city's has no **ligthing** for months .

Vectors for download

http://mumin.ling.su.se/fotran2018/

Can they be used for NMT transfer?

- Many-to-English system with fixed multilingual embeddings at encoder
- 10M sentences sampled from WMT news task: Czech, Russian, Turkish, Finnish, Estonian
- Enough to learn a good decoder-side (English) LM, reasonable translation model for the training domain
- But can it translate Bible text...?

"Zero-shot" transfer... needs some work still

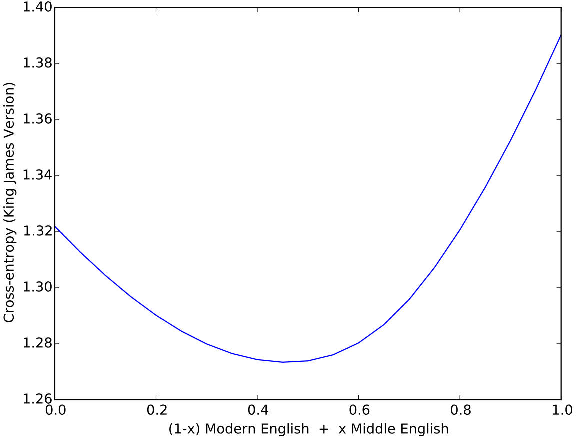
Source	Nam di lo: nay ma ŋra mulda vi Lawna hidi mige? Lebo nay as ŋra' hidi law ma nir-niramna mige?
Translation	he said again, "We were the kingdom of God and what or we do not compare and speak."
Reference	And he said, "With what can we compare the kingdom of God, or what parable shall we use for it?

- Massa [mcn], about 200 000 speakers in Chad and Cameroon
- Overly optimistic result, since this sentence was used in projecting the embeddings

Representations

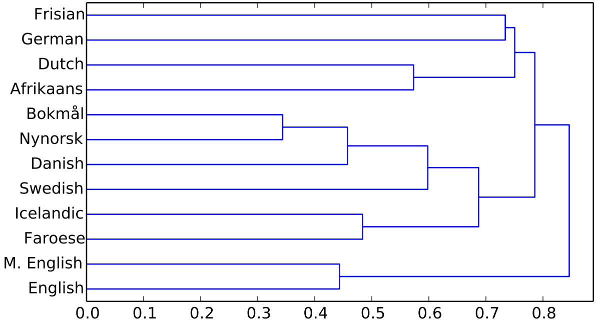
- Multilingual word embeddings encode vocabularies
- How to encode "grammar" in a highly multilingual model?
- We can condition a neural model on the language used for each example

Language representations



Proof-of-concept with language modeling

Language representations



Structure in (part of) the language space discovered

Future work

- To what extent can we frame universal MT as...
  1. multilingual word representations (lexicon)
  2. language representations (grammar)
  3. neural model (the machine)
- How to model the strong cross-lingual patterns of grammar and lexicon?
- How to integrate diverse information sources from linguistics?
- ...or should we just write rules and/or create more data?