THE UNIVERSITY of EDINBURGH

# Machine Translation

## 11: Multilingual and Zero-Shot Translation

Rico Sennrich
(slide credit: Adam Lopez)

University of Edinburgh

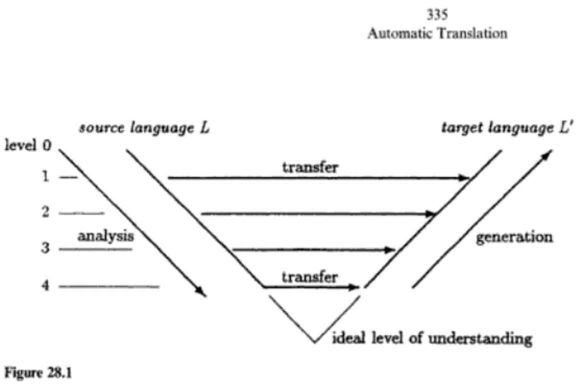


# Multilingual translation

- Isn't translation already multilingual?
- Consider these datasets:
  - United Nations (6 languages)
  - European parliament (21 languages)
  - The Bible (484 complete and 2551 partial translations)
- What can we do with more than two languages?



# Interlingua




from Vauquois, 1968



# Multi-source translation

# Multi-source translation

- Suppose we have a document in French and its (human) translation in German. Questions:



# Multi-source translation

- Suppose we have a document in French and its (human) translation in German. Questions:
  - Will it help to use both translations?



# Multi-source translation

- Suppose we have a document in French and its (human) translation in German. Questions:
  - Will it help to use both translations?
  - Is there any use for this?

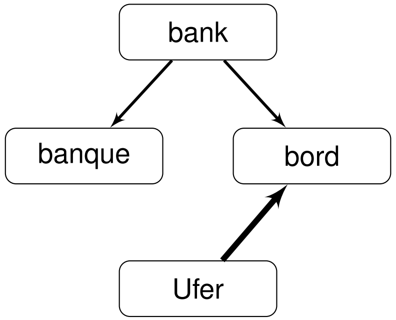


# Multi-source translation

- Suppose we have a document in French and its (human) translation in German. Questions:
  - Will it help to use both translations?
  - Is there any use for this?
    - YES! The European Parliament has 24 official languages. But not all translation is directly from a source and target language; there is often a bridge language.

## Motivation for Multi-Source Translation

ambiguities in one source language may be resolved in the other and vice versa






# Multi-source translation

Quite an old idea (e.g. Och & Ney 2001)

Table 4: Absolute improvements in WER combining two languages using method MAX compared with the best WER obtained by any of the two languages.

| | fr | pt | es | it | sv | da | nl |
|---|---|---|---|---|---|---|---|
| fr | 0.0 | 1.5 | 1.2 | 0.5 | 2.7 | 1.9 | 0.8 |
| pt | | 0.0 | 2.2 | 2.1 | 4.0 | 3.4 | 1.3 |
| es | | | 0.0 | 2.4 | 3.9 | 2.6 | 1.7 |
| it | | | | 0.0 | 3.5 | 3.2 | 1.6 |
| sv | | | | | 0.0 | 2.7 | 1.7 |
| da | | | | | | 0.0 | 4.3 |
| nl | | | | | | | 0.0 |

Table 5: Absolute improvements in WER combining two languages using method PROD compared with the best WER obtained by any of the two languages.

| | fr | pt | es | it | sv | da | nl |
|---|---|---|---|---|---|---|---|
| fr | 0.0 | 0.8 | 0.1 | 0.4 | 1.0 | 0.8 | -0.2 |
| pt | | 0.0 | 2.6 | 2.1 | 2.6 | 2.8 | -0.1 |
| es | | | 0.0 | 2.4 | 3.4 | 3.7 | 1.1 |
| it | | | | 0.0 | 1.9 | 3.0 | 0.3 |
| sv | | | | | 0.0 | 1.8 | 0.5 |
| da | | | | | | 0.0 | 1.5 |
| nl | | | | | | | 0.0 |

Table 6: Language combination using method MAX.

| languages | WER | PER |
|---|---|---|
| fr | 55.3 | 45.3 |
| fr+sv | 52.6 | 43.7 |
| fr+sv+es | **52.0** | **43.2** |
| fr+sv+es+pt | 52.3 | 43.6 |
| fr+sv+es+pt+it | 52.7 | 44.0 |
| fr+sv+es+pt+it+da | 52.5 | 43.9 |

Table 7: Language combination using method PROD.

| languages | WER | PER |
|---|---|---|
| fr | 55.3 | 45.3 |
| fr+sv | 54.3 | 44.5 |
| fr+sv+es | 51.0 | 41.4 |
| fr+sv+es+pt | 50.2 | 40.2 |
| fr+sv+es+pt+it | 49.8 | 39.8 |
| fr+sv+es+pt+it+da | **48.8** | **39.1** |



# Multi-source translation



# Multi-source translation

- Assorted techniques to do this in IBM-style or phrase-based MT.

# Multi-source translation

- Assorted techniques to do this in IBM-style or phrase-based MT.
- Difficult to model directly due to independence assumptions of these models.



# Multi-source translation

- Assorted techniques to do this in IBM-style or phrase-based MT.
- Difficult to model directly due to independence assumptions of these models.
- Usually done as a kind of system combination (merging the output of two MT systems).



# Multi-source translation

- Assorted techniques to do this in IBM-style or phrase-based MT.
- Difficult to model directly due to independence assumptions of these models.
- Usually done as a kind of system combination (merging the output of two MT systems).
- But this introduces other problems, e.g. decoding.



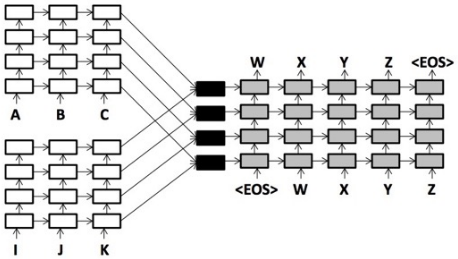
# Multi-source translation

- Assorted techniques to do this in IBM-style or phrase-based MT.
- Difficult to model directly due to independence assumptions of these models.
- Usually done as a kind of system combination (merging the output of two MT systems).
- But this introduces other problems, e.g. decoding.
- Fundamentally, it’s interpolation of conditional LMs.

# Direct multi-source

Zoph & Knight 2016

- Directly learns and uses *p(English|French,German)*
- For attention: two context vectors (uses p-local attention of Luong, et al, but could use other methods).

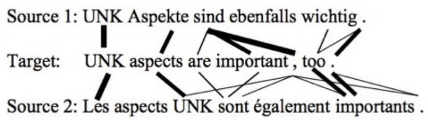





# Direct multi-source

Zoph & Knight 2016

- Directly learns and uses *p(English|French,German)*
- For attention: two context vectors (uses p-local attention of Luong, et al, but could use other methods).




| Target = English | | | |
|---|---|---|---|
| **Source** | **Method** | **Ppl** | **BLEU** |
| French | — | 10.3 | 21.0 |
| German | — | 15.9 | 17.3 |
| French+German | Basic | 8.7 | 23.2 |
| French+German | Child-Sum | 9.0 | 22.5 |
| French+French | Child-Sum | 10.9 | 20.7 |
| French | Attention | 8.1 | 25.2 |
| French+German | B-Attent. | 5.7 | 30.0 |
| French+German | CS-Attent. | 6.0 | 29.6 |



# Multi-way MT

Firat et al. 2016 (two papers)



# Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.

# Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.



# Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.
- But what about attention?



# Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.
- But what about attention?

$$p(f_i|f_{i-1},...,f_1,\mathbf{e}) = g(f_{i-1}, s_i, c_i)$$

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j \qquad \alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|}\exp(a_{ik})}$$

$$a_{ij} = a(s_{i-1}, h_j)$$



# Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.
- But what about attention?

$$p(f_i|f_{i-1},...,f_1,\mathbf{e}) = g(\boxed{f_{i-1}, s_i, c_i})$$

Everything we need is right here!

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j \qquad \alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|}\exp(a_{ik})}$$

$$a_{ij} = a(s_{i-1}, h_j)$$

# Multi-way MT

Firat et al. 2016 (two papers)

- As in Bahdanu et al. (2014), attention mechanism is a feedforward function of both decoder hidden state and encoder context vector.
- **Shared** between all encoders and decoders.

$$p(f_i|f_{i-1}, ..., f_1, \mathbf{e}) = g(\boxed{f_{i-1}, s_i, c_i})$$

Everything we need is right here!

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j \qquad \alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|} \exp(a_{ik})}$$

$$a_{ij} = a(s_{i-1}, h_j)$$



# Multi-way MT

Firat et al. 2016 (two papers)

| | Size | Single | Single+DF | Multi |
|---|---|---|---|---|
| En→Fi | 100k | 5.06/3.96 | 4.98/3.99 | 6.2/**5.17** |
| | 200k | 7.1/6.16 | 7.21/6.17 | 8.84/**7.53** |
| | 400k | 9.11/7.85 | 9.31/8.18 | 11.09/**9.98** |
| | 800k | 11.08/9.96 | 11.59/10.15 | 12.73/**11.28** |
| De→En | 210k | 14.27/13.2 | 14.65/13.88 | 16.96/**16.26** |
| | 420k | 18.32/17.32 | 18.51/17.62 | 19.81/**19.63** |
| | 840k | 21/19.93 | 21.69/20.75 | 22.17/**21.93** |
| | 1.68m | 23.38/23.01 | 23.33/22.86 | 23.86/**23.52** |
| En→De | 210k | 11.44/11.57 | 11.71/11.16 | 12.63/**12.68** |
| | 420k | 14.28/14.25 | 14.88/15.05 | 15.01/**15.67** |
| | 840k | 17.09/17.44 | 17.21/17.88 | 17.33/**18.14** |
| | 1.68m | 19.09/19.6 | 19.36/20.13 | 19.23/**20.59** |

**Table 2:** BLEU scores where the target pair's parallel corpus is constrained to be 5%, 10%, 20% and 40% of the original size. We report the BLEU scores on the development and test sets (separated by /) by the single-pair model (Single), the single-pair model with monolingual corpus (Single+DF) and the proposed multi-way, multilingual model (Multi).

Low-resource **simulation** (using high-resource European languages)



# Multi-way MT

Firat et al. 2016 (two papers)

| | | Dir | Fr (39m) → En | Fr (39m) En → | Cs (12m) → En | Cs (12m) En → | De (4.2m) → En | De (4.2m) En → | Ru (2.3m) → En | Ru (2.3m) En → | Fi (2m) → En | Fi (2m) En → |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) BLEU | Dev | Single | 27.22 | 26.91 | 21.24 | 15.9 | 24.13 | 20.49 | 21.04 | 18.06 | 13.15 | 9.59 |
| | | Multi | 26.09 | 25.04 | 21.23 | 14.42 | 23.66 | 19.17 | 21.48 | 17.89 | 12.97 | 8.92 |
| | Test | Single | 27.94 | **29.7** | 20.32 | **13.84** | 24 | **21.75** | 22.44 | **19.54** | 12.24 | **9.23** |
| | | Multi | **28.06** | 27.88 | **20.57** | 13.29 | **24.20** | 20.59 | **23.44** | 19.39 | **12.61** | 8.98 |
| (b) LL | Dev | Single | -50.53 | -53.38 | -60.69 | -69.56 | -54.76 | -61.21 | -60.19 | -65.81 | -88.44 | -91.75 |
| | | Multi | -50.6 | -56.55 | -54.46 | -70.76 | -54.14 | -62.34 | -54.09 | -63.75 | -74.84 | -88.02 |
| | Test | Single | -43.34 | **-45.07** | -60.03 | **-64.34** | -57.81 | **-59.55** | -60.65 | -60.29 | -88.66 | -94.23 |
| | | Multi | **-42.22** | -46.29 | **-54.66** | -64.80 | **-53.85** | -60.23 | **-54.49** | **-58.63** | **-71.26** | **-88.09** |

**Table 3:** (a) BLEU scores and (b) average log-probabilities for all the five languages from WMT'15.



# Multi-way MT

Firat et al. 2016 (two papers)

| | | Dir | Fr (39m) → En | Fr (39m) En → | Cs (12m) → En | Cs (12m) En → | De (4.2m) → En | De (4.2m) En → | Ru (2.3m) → En | Ru (2.3m) En → | Fi (2m) → En | Fi (2m) En → |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) BLEU | Dev | Single | 27.22 | 26.91 | 21.24 | 15.9 | 24.13 | 20.49 | 21.04 | 18.06 | 13.15 | 9.59 |
| | | Multi | 26.09 | 25.04 | 21.23 | 14.42 | 23.66 | 19.17 | 21.48 | 17.89 | 12.97 | 8.92 |
| | Test | Single | 27.94 | **29.7** | 20.32 | **13.84** | 24 | **21.75** | 22.44 | **19.54** | 12.24 | **9.23** |
| | | Multi | **28.06** | 27.88 | **20.57** | 13.29 | **24.20** | 20.59 | **23.44** | 19.39 | **12.61** | 8.98 |
| (b) LL | Dev | Single | -50.53 | -53.38 | -60.69 | -69.56 | -54.76 | -61.21 | -60.19 | -65.81 | -88.44 | -91.75 |
| | | Multi | -50.6 | -56.55 | -54.46 | -70.76 | -54.14 | -62.34 | -54.09 | -63.75 | -74.84 | -88.02 |
| | Test | Single | -43.34 | **-45.07** | -60.03 | **-64.34** | -57.81 | **-59.55** | -60.65 | -60.29 | -88.66 | -94.23 |
| | | Multi | **-42.22** | -46.29 | **-54.66** | -64.80 | **-53.85** | -60.23 | **-54.49** | **-58.63** | **-71.26** | **-88.09** |

**Table 3:** (a) BLEU scores and (b) average log-probabilities for all the five languages from WMT'15.

ok, but what about multi-source?

# Multi-way multi-source MT

Firat et al. 2016 (two papers)



# Multi-way multi-source MT

Firat et al. 2016 (two papers)

- Still assumes only many bilingual parallel corpora.
- What to do if there are multiple input sentences?
- Early averaging (average context vectors).
- Late averaging (aka linear interpolation).



# Multi-way multi-source MT

Firat et al. 2016 (two papers)

- Still assumes only many bilingual parallel corpora.
- What to do if there are multiple input sentences?
- Early averaging (average context vectors). $\mathbf{c}_t = \frac{\mathbf{c}_t^1 + \mathbf{c}_t^2}{2}.$
- Late averaging (aka linear interpolation).

$$P(w_i|\boldsymbol{c}) = \sum_{k=1}^{K} \lambda_k(\boldsymbol{c}) P_k(w_i|\boldsymbol{c})$$



# Multi-way multi-source MT

Firat et al. 2016 (two papers)

- Still assumes only many bilingual parallel corpora.
- What to do if there are multiple input sentences?
- Early averaging (average context vectors). $\mathbf{c}_t = \frac{\mathbf{c}_t^1 + \mathbf{c}_t^2}{2}.$
- Late averaging (aka linear interpolation).

$$P(w_i|\boldsymbol{c}) = \sum_{k=1}^{K} \lambda_k(\boldsymbol{c}) P_k(w_i|\boldsymbol{c})$$

Early and late averaging are orthogonal, can be combined.

# Multi-way multi-source MT

Firat et al. 2016 (two papers)

| | Src | Trgt | Multi Dev | Multi Test | Single Dev | Single Test |
|---|---|---|---|---|---|---|
| (a) | Es | En | 30.73 | 28.32 | 29.74 | 27.48 |
| (b) | Fr | En | 26.93 | 27.93 | 26.00 | 27.21 |
| (c) | En | Es | 30.63 | 28.41 | 31.31 | 28.90 |
| (d) | En | Fr | 22.68 | 23.41 | 22.80 | 24.05 |

**Table 2:** One-to-one translation qualities using the multi-way, multilingual model and four separate single-pair models.



# Multi-way multi-source MT

Firat et al. 2016 (two papers)

| | Src | Trgt | Multi Test | Single Test |
|---|---|---|---|---|
| (a) | Es | En | 28.32 | 27.48 |
| (b) | Fr | En | 27.93 | 27.21 |
| (c) | En | Es | 28.41 | 28.90 |
| (d) | En | Fr | 23.41 | 24.05 |

**Table 2:** One-to-one translation qualities using the multi-way, multilingual model and four separate single-pair models.



# Multi-way multi-source MT

Firat et al. 2016 (two papers)

| | Src | Trgt | Multi Test | Single Test |
|---|---|---|---|---|
| (a) | Es | En | 28.32 | 27.48 |
| (b) | Fr | En | 27.93 | 27.21 |

**Table 2:** One-to-one translation qualities using the multi-way multilingual model and four separate single-pair models.

| | | Multi Dev | Multi Test | Single Dev | Single Test |
|---|---|---|---|---|---|
| (a) | Early | 31.89 | 31.35 | – | – |
| (b) | Late | 32.04 | 31.57 | 32.00 | 31.46 |
| (c) | E+L | 32.61 | 31.88 | – | – |

**Table 3:** Many-to-one quality (Es+Fr→En) using three translation strategies. Compared to Table 2 (a–b) we observe a significant improvement (up to 3+ BLEU), although the model was never trained in these many-to-one settings. The second column shows the quality by the ensemble of two separate single-pair models.



# Multi-way multi-source MT

Firat et al. 2016 (two papers)

| | Src | Trgt | Multi Test | Single Test |
|---|---|---|---|---|
| (a) | Es | En | 28.32 | 27.48 |
| (b) | Fr | En | 27.93 | 27.21 |

**Table 2:** One-to-one translation qualities using the multi-way multilingual model and four separate single-pair models.

| | | Multi Dev | Multi Test | Single Dev | Single Test |
|---|---|---|---|---|---|
| (a) | Early | 31.89 | 31.35 | – | – |
| (b) | Late | 32.04 | 31.57 | 32.00 | 31.46 |
| (c) | E+L | 32.61 | 31.88 | – | – |

**Table 3:** Many-to-one quality (Es+Fr→En) using three translation strategies. Compared to Table 2 (a–b) we observe a significant improvement (up to 3+ BLEU), although the model was never trained in these many-to-one settings. The second column shows the quality by the ensemble of two separate single-pair models.

# Zero-shot MT

Firat et al. 2016 (two papers)

- Suppose our bilingual parallel data include a pair of languages for which we have no parallel data.

Spanish ⟷ English    English ⟷ French

- Q: Can we use the multi-way encoder-decoder system to translate Spanish into French?



# Zero-shot MT

Firat et al. 2016 (two papers)

- Suppose our bilingual parallel data include a pair of languages for which we have no parallel data.
- Q: Can we use the multi-way encoder-decoder system to translate Spanish into French?

| | Pivot | Many-to-1 | Dev | Test |
|---|---|---|---|---|
| (a) | | | < 1 | < 1 |
| (b) | √ | | 20.64 | 20.4 |

**Table 4:** Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning. When pivot is √, English is used as a pivot language.

A: Not really

Must pivot (explicitly) through English



# Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data?

| | Pivot | Many-to-1 | Dev | Test |
|---|---|---|---|---|
| (a) | | | < 1 | < 1 |
| (b) | √ | | 20.64 | 20.4 |

**Table 4:** Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning. When pivot is √, English is used as a pivot language.



# Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**

Spanish ⟷ English    English ⟷ French

# Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**

Spanish ⟷ English    English ⟷ French

↕

Spanish (MT)



# Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**

Spanish ⟷ English    English ⟷ French

↕

Spanish (MT)



# Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?

| | | | Pseudo Parallel Corpus | | | |
|---|---|---|---|---|---|---|
| Pivot | Many-to-1 | | 1k | 10k | 100k | 1m |
| Single-Pair Models | | Dev | – | – | – | – |
| | | Test | – | – | – | – |
| √ | No Finetuning | | Dev: 20.64, Test: 20.4 | | | |
| | | Dev | 0.28 | 10.16 | 15.61 | 17.59 |
| | | Test | 0.47 | 10.14 | 15.41 | 17.61 |



# Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?

| | | | Pseudo Parallel Corpus | | | | True Parallel Corpus | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pivot | Many-to-1 | | 1k | 10k | 100k | 1m | 1k | 10k | 100k | 1m |
| Single-Pair Models | | Dev | – | – | – | – | – | – | 11.25 | 21.32 |
| | | Test | – | – | – | – | – | – | 10.43 | 20.35 |
| √ | No Finetuning | | Dev: 20.64, Test: 20.4 | | | | – | | | |
| | | Dev | 0.28 | 10.16 | 15.61 | 17.59 | 0.1 | 8.45 | 16.2 | 20.59 |
| | | Test | 0.47 | 10.14 | 15.41 | 17.61 | 0.12 | 8.18 | 15.8 | 19.97 |

# Zero-shot MT

Johnson et al. 2016 (Google)

- Do we really need *N* encoders and *N* decoders?
- Can we just learn a single function parameterized by the desired output language?
  - Implementation: add a token indicating desired output language to input.
- Why is this a nice solution (for Google)?



# Multi-source MT

Johnson et al. 2016 (Google)

- Sanity check: must not make things worse.

Table 1: Many to One: BLEU scores on various data sets for single language pair and multilingual models.

| Model | Single | Multi | Diff |
|---|---|---|---|
| WMT German→English (oversampling) | 30.43 | 30.59 | +0.16 |
| WMT French→English (oversampling) | 35.50 | 35.73 | +0.23 |
| WMT German→English (no oversampling) | 30.43 | 30.54 | +0.11 |
| WMT French→English (no oversampling) | 35.50 | 36.77 | +0.27 |
| Prod Japanese→English | 23.41 | 23.87 | +0.46 |
| Prod Korean→English | 25.42 | 25.47 | +0.05 |
| Prod Spanish→English | 38.00 | 38.73 | +0.73 |
| Prod Portuguese→English | 44.40 | 45.19 | +0.79 |



# Multi-*target* MT

Johnson et al. 2016 (Google)

- Sanity check: must not make things worse.

Table 2: One to Many: BLEU scores on various data sets for single language pair and multilingual models.

| Model | Single | Multi | Diff |
|---|---|---|---|
| WMT English→German (oversampling) | 24.67 | 24.97 | +0.30 |
| WMT English→French (oversampling) | 38.95 | 36.84 | -2.11 |
| WMT English→German (no oversampling) | 24.67 | 22.61 | -2.06 |
| WMT English→French (no oversampling) | 38.95 | 38.16 | -0.79 |
| Prod English→Japanese | 23.66 | 23.73 | +0.07 |
| Prod English→Korean | 19.75 | 19.58 | -0.17 |
| Prod English→Spanish | 34.50 | 35.40 | +0.90 |
| Prod English→Portuguese | 38.40 | 38.63 | +0.23 |



# Zero-shot MT

Johnson et al. 2016 (Google)

- Incremental training: add a small amount of (true) parallel data in the language pair of interest.

Table 5: Portuguese→Spanish BLEU scores using various models.

| | Model | BLEU |
|---|---|---|
| (a) | PBMT bridged | 28.99 |
| (b) | NMT bridged | 30.91 |
| (c) | NMT Pt→Es | 31.50 |
| (d) | Model 1 (Pt→En, En→Es) | 21.62 |
| (e) | Model 2 (En↔{Es, Pt}) | 24.75 |
| (f) | Model 2 + incremental training | 31.77 |

# Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

| | Zero-Shot | From-Scratch | Incremental |
|---|---|---|---|
| English→Belarusian | 16.85 | 17.03 | 16.99 |
| English→Russian | 22.21 | 22.03 | 21.92 |
| English→Ukrainian | 18.16 | 17.75 | 18.27 |
| Belarusian→English | 25.44 | 24.72 | 25.54 |
| Russian→English | 28.36 | 27.90 | 28.46 |
| Ukrainian→English | 28.60 | 28.51 | 28.58 |
| Belarusian→Russian | 56.53 | 82.50 | 78.63 |
| Russian→Belarusian | 58.75 | 72.06 | 70.01 |
| Russian→Ukrainian | 21.92 | 25.75 | 25.34 |
| Ukrainian→Russian | 16.73 | 30.53 | 29.92 |

trained on parallel data




# Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

| | Zero-Shot | From-Scratch | Incremental |
|---|---|---|---|
| English→Belarusian | 16.85 | 17.03 | 16.99 |
| English→Russian | 22.21 | 22.03 | 21.92 |
| English→Ukrainian | 18.16 | 17.75 | 18.27 |
| Belarusian→English | 25.44 | 24.72 | 25.54 |
| Russian→English | 28.36 | 27.90 | 28.46 |
| Ukrainian→English | 28.60 | 28.51 | 28.58 |
| Belarusian→Russian | 56.53 | 82.50 | 78.63 |
| Russian→Belarusian | 58.75 | 72.06 | 70.01 |
| Russian→Ukrainian | 21.92 | 25.75 | 25.34 |
| Ukrainian→Russian | 16.73 | 30.53 | 29.92 |


zero-shot + small parallel data




# Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

| | Zero-Shot | From-Scratch | Incremental |
|---|---|---|---|
| English→Belarusian | 16.85 | 17.03 | 16.99 |
| English→Russian | 22.21 | 22.03 | 21.92 |
| English→Ukrainian | 18.16 | 17.75 | 18.27 |
| Belarusian→English | 25.44 | 24.72 | 25.54 |
| Russian→English | 28.36 | 27.90 | 28.46 |
| Ukrainian→English | 28.60 | 28.51 | 28.58 |
| Belarusian→Russian | 56.53 | 82.50 | 78.63 |
| Russian→Belarusian | 58.75 | 72.06 | 70.01 |
| Russian→Ukrainian | 21.92 | 25.75 | 25.34 |
| Ukrainian→Russian | 16.73 | 30.53 | 29.92 |


actual zero-shot experiment



# Zero-shot MT

Johnson et al. 2016 (Google)

code-switching in the input language:

**Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.

**Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.

**Mixed Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

code-switching in the output language:

| Spanish/Portuguese: | Here the other guinea-pig cheered, and was suppressed. |
|---|---|
| $w_{pt} = 0.00$ | Aquí el otro conejillo de indias animó, y fue suprimido. |
| $w_{pt} = 0.30$ | Aquí el otro conejillo de indias animó, y fue suprimido. |
| $w_{pt} = 0.40$ | Aquí, o outro porquinho-da-índia alegrou, e foi suprimido. |
| $w_{pt} = 0.42$ | Aqui o outro porquinho-da-índia alegrou, e foi suprimido. |
| $w_{pt} = 0.70$ | Aqui o outro porquinho-da-índia alegrou, e foi suprimido. |
| $w_{pt} = 0.80$ | Aqui a outra cobaia animou, e foi suprimida. |
| $w_{pt} = 1.00$ | Aqui a outra cobaia animou, e foi suprimida. |

# Zero-shot MT

Johnson et al. 2016 (Google)

Portuguese informant: “we decided it's impossible to judge the correctness of the translation without context (but it's likely wrong). After finding the context (Alice in Wonderland) we can conclude it's wrong.”

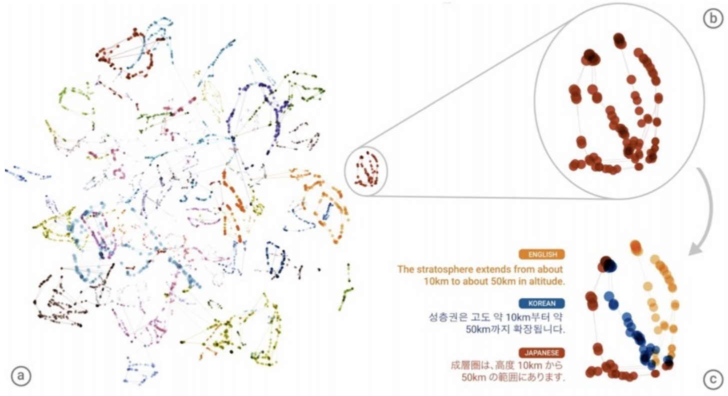
code-switching in the output language:

| Spanish/Portuguese: | Here the other guinea-pig cheered, and was suppressed. |
|---|---|
| $w_{pt} = 0.00$ | Aquí el otro conejillo de indias animó, y fue suprimido. |
| $w_{pt} = 0.30$ | Aquí el otro conejillo de indias animó, y fue suprimido. |
| $w_{pt} = 0.40$ | Aquí, o outro porquinho-da-índia alegrou, e foi suprimido. |
| $w_{pt} = 0.42$ | Aqui o outro porquinho-da-índia alegrou, e foi suprimido. |
| $w_{pt} = 0.70$ | Aqui o outro porquinho-da-índia alegrou, e foi suprimido. |
| $w_{pt} = 0.80$ | Aqui a outra cobaia animou, e foi suprimida. |
| $w_{pt} = 1.00$ | Aqui a outra cobaia animou, e foi suprimida. |



# Zero-shot MT

Johnson et al. 2016 (Google)

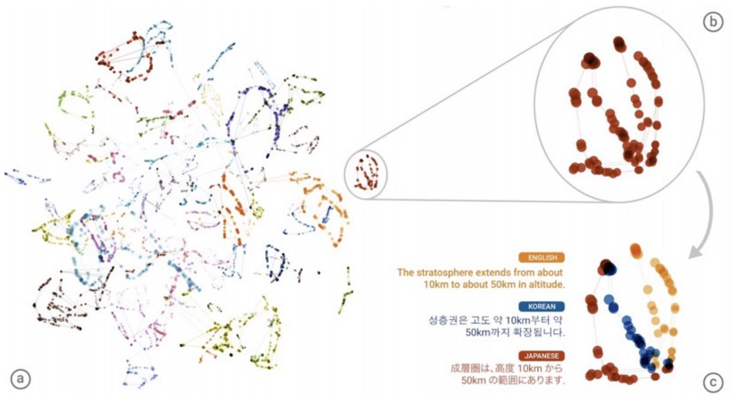



Low-dimensional embeddings of context vectors



# Zero-shot MT

Johnson et al. 2016 (Google)




Low-dimensional embeddings of context vectors

Provocative (untestable) claim: this is an interlingua