On Various Approaches to Machine Translation from Russian to Kazakh

Aibek Makazhanov[a], Bagdat Myrzakhmetov[a,b], Zhanibek Kozhirbayev[a,c]

(a) National Laboratory Astana, Nazarbayev University, Astana
(b) School of Science and Technology, Nazarbayev University, Astana
(c) L.N. Gumilyov Eurasian National University, Astana

Outline

Motivation: Why Russian to Kazakh?

- **Strong demand**: someone is probably doing R2K translation as I speak;

Motivation: Why Russian to Kazakh?

- **Strong demand**: someone is probably doing R2K translation as I speak;
- **Resource rich**: tons of accessible comparable text;

Motivation: Why Russian to Kazakh?

- **Strong demand**: someone is probably doing R2K translation as I speak;
- **Resource rich**: tons of accessible comparable text;
- **Research poor**: not too many people are doing it;

Motivation: Why Russian to Kazakh?

- **Strong demand**: someone is probably doing R2K translation as I speak;
- **Resource rich**: tons of accessible comparable text;
- **Research poor**: not too many people are doing it;
- **Interesting problem**: both languages are `MCL`.

Approaches to MT: Linguistically motivated (RBMT)

You will need:

- linguistic proficiency;
- lexicons for source and target language;
- tools: analyzers, taggers, parsers;
- transfer rules.

Approaches to MT: Linguistically motivated (RBMT)

You will need:

- linguistic proficiency;
- lexicons for source and target language;
- tools: analyzers, taggers, parsers;
- transfer rules.

tl;dr: linguists + pain & suffering (tons of) = result.

Approaches to MT: Data-driven (NMT)

You will need:

- Encoder: source sentence $\rightarrow$ iVector;
- iVector $\rightarrow$ Decoder (MAGIC) $\rightarrow$ target sentence;
- parallel corpus;
- additional monolingual corpus for target L (desirable);
- GPUs.

Approaches to MT: Data-driven (NMT)

You will need:

- Encoder: source sentence $\rightarrow$ iVector;
- iVector $\rightarrow$ Decoder (MAGIC) $\rightarrow$ target sentence;
- parallel corpus;
- additional monolingual corpus for target L (desirable);
- GPUs.

tl;dr: data $+$ DL framework $=$ result.

Approaches to MT: Data-driven (SMT)

You will need:

- $P(t|s) = P(s|t)P(t)$;
- parallel corpus;
- additional monolingual corpus for target L (desirable);
- decoder.

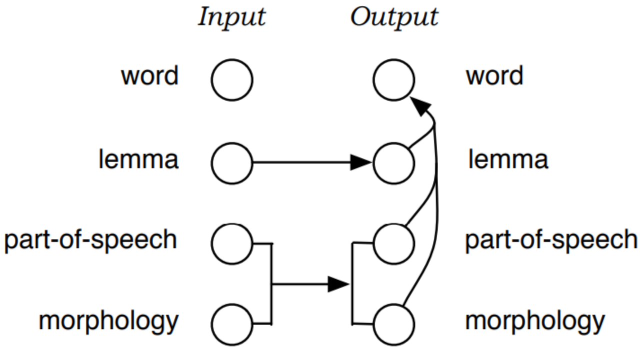
Approaches to MT: Data-driven (SMT)

You will need:

- $P(t|s) = P(s|t)P(t)$;
- parallel corpus;
- additional monolingual corpus for target L (desirable);
- decoder.

tl;dr: data + Moses stack = result.

Approaches to MT: Hybrid (Factored SMT)



Approaches to MT: Hybrid (Factored SMT)

You will need:

- $P(t|s) = P(s|t)P(t)$;
- (annotated!) parallel corpus;
- additional (annotated!) monolingual corpus for target L;
- decoder.

Approaches to MT: Hybrid (Factored SMT)

You will need:

- $P(t|s) = P(s|t)P(t)$;
- (annotated!) parallel corpus;
- additional (annotated!) monolingual corpus for target L;
- decoder.

tl;dr: data + pain & suffering + Moses stack = result.

Approaches to MT

- **Linguistically motivated**: RBMT
  - (+) easy to control/debug;
  - (−) hard to develop/adapt;
- **Data-driven**: SMT, NMT
  - (+) relatively easy to develop/adapt;
  - (−) hard to control/debug;
- **Hybrid (Data + Linguistics)**: factored SMT
  - (+) supposed to be great for MCL;
  - (−) hard to develop/adapt/control/debug.

Parallel Corpus

Website	Aligned	Cleaned	Filtered	Test set	Tuning set
www.akorda.kz	80 333	75 240	75 199	150	100
www.primeminister.kz	111 193	81 060	79 483	140	95
www.ortcom.kz	77 468	73 770	73 610	130	100
www.nurotan.kz	63 268	57 043	56 563	90	60
www.astana.gov.kz	105 929	91 010	90 762	150	100
www.strategy2050.kz	372 249	347 560	345 372	495	310
www.adilet.gov.kz	59 489	30 083	28 744	75	50
www.economy.gov.kz	15 179	11 417	11 398	20	15
www.dkz.mzsr.gov.kz	24 813	7 624	7 232	15	15
www.kaztag.kz	52 643	45 653	45 505	75	50
www.almaty.gov.kz	25 176	18 211	18 036	30	20
www.mfa.gov.kz	12 463	10 466	10 405	20	10
www.palata.kz	36 394	33 355	33 206	70	50
www.expo2017astana.com	7 244	6 027	5 963	10	5
www.emer.gov.kz	14 199	11 855	11 756	30	20
Total	**1 069 078**	**909 189**	**893 234**	**1 500**	**1 000**

Parallel Corpus: Alignment

- Normalization: (> = $>$) || (cepi = *cepi*);
- Sentence splitting: **Punkt** (Kiss and Strunk, 2006);
- Lemmatization: **MyStem** (Segalovich, 2003);
- Sentence alignment: **Hunalign** (Varga et al., 2007).

Parallel Corpus: Cleaning

Remove the following aligned sentence pairs:

- Duplicates;
- $S1 = S2$;
- No alphabetics in either;
- $L < 3$ and $L > 50$.

Parallel Corpus: Filtering

Train and apply ML classifier that uses these features:

#	DC	Feature	---------	#	DC	Feature
1,2	S,T	length in characters		19,20	S,T	count of personal initials
3	ST	MMR(F1,F2)		21	ST	COS(F19*,F20*)
4,5	S,T	length in tokens		22,23	S,T	ratio of alphanumerics
6	ST	MMR(F4,F5)		24	ST	MMR(F22,F23)
7,8	S,T	count of symbols		25,26	S,T	count of words in quotes
9	ST	COS(F7*,F8*)		27	ST	MMR(F25,F26)
10,11	S,T	count of numerals		28,29	S,T	count of words in parenthesis
12	ST	COS(F10*,F11*)		30	ST	MMR(F25,F26)
13,14	S,T	count of digits		31	ST	num. of tokens in identical pairs
15	ST	COS(F13*,F14*)		32	ST	min-max ratio between unique tokens in source and target sentences
16,17	S,T	count of latin alphanumerics				
18	ST	COS(F16*,F17*)		33-35	ST	`Hunalign` score: absolute, relative, min-max scaled.

Parallel Corpus

Website	Aligned	Cleaned	Filtered	Test set	Tuning set
www.akorda.kz	80 333	75 240	75 199	150	100
www.primeminister.kz	111 193	81 060	79 483	140	95
www.ortcom.kz	77 468	73 770	73 610	130	100
www.nurotan.kz	63 268	57 043	56 563	90	60
www.astana.gov.kz	105 929	91 010	90 762	150	100
www.strategy2050.kz	372 249	347 560	345 372	495	310
www.adilet.gov.kz	59 489	30 083	28 744	75	50
www.economy.gov.kz	15 179	11 417	11 398	20	15
www.dkz.mzsr.gov.kz	24 813	7 624	7 232	15	15
www.kaztag.kz	52 643	45 653	45 505	75	50
www.almaty.gov.kz	25 176	18 211	18 036	30	20
www.mfa.gov.kz	12 463	10 466	10 405	20	10
www.palata.kz	36 394	33 355	33 206	70	50
www.expo2017astana.com	7 244	6 027	5 963	10	5
www.emer.gov.kz	14 199	11 855	11 756	30	20
Total	**1 069 078**	**909 189**	**893 234**	**1 500**	**1 000**

Comparing data-driven approaches: setup

- **SMT**:
  - Moses + KenLM;
  - trigram LMs;
  - MERT tuning;
- **NMT**:
  - (i) LSTM, (ii) +Attention 2, (iii) +Attention 4;
  - vocabulary – 50K most frequent;
  - 128 hidden units;
  - 0.2 dropout probability;
  - 20000 iterations on Tesla k2025 GPU;
- **Evaluation**: automatic (BLEU)

Comparing data-driven approaches: results

Model	BLEU
SMT	**34.15**
Basic NMT	3.90
Attention NMT, 2 layers	9.14
Attention NMT, 4 layers	11.00

Comparing all of the approaches: setup

- **RBMT**:
  - Apertium (rev. 82385);
  - experimental system;
- **(factored) SMT**:
  - factors: lemma, POS;
  - 1/5 training set;
- **Evaluation**: automatic (BLEU)

Comparing data-driven approaches: results

Model	**BLEU**
RBMT	6.41
Factored SMT	21.77
SMT	**24.73**

Conclusions and Future Work

We have

- assembled a parallel R2K corpus;
- for R2K compared most popular MT approaches;

Conclusions and Future Work

We have

- assembled a parallel R2K corpus;
- for R2K compared most popular MT approaches;

We now know that

- NMT needs more resources;
- factored SMT needs more than lemma+POS;

Conclusions and Future Work

We have

- assembled a parallel R2K corpus;
- for R2K compared most popular MT approaches;

We now know that

- NMT needs more resources;
- factored SMT needs more than lemma+POS;

We will

- expand the parallel corpus;
- look for better NMT representations;
- experiment with various factors.