NPFL120 Multilingual Natural Language Processing

Machine Translation: Alignment and Phrase-Based MT

Ondřej Bojar

March 11, 2019

LANGTECH

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline of Lectures on MT

Today:

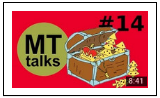
- Document, sentence and word alignment.
- Phrase-based MT as a tool.

Towards the end of semester:

- Multi-lingual (neural) MT.

Supplementary Materials

Videolectures & Wiki:



http://mttalks.ufal.ms.mff.cuni.cz/

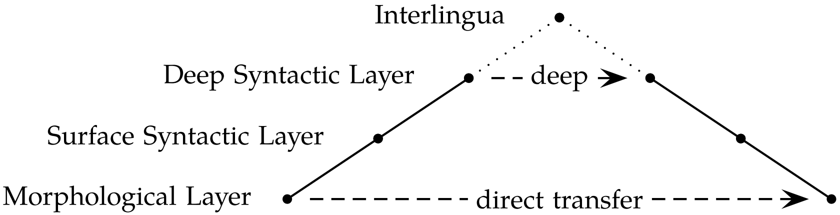
Slides and Lectures from MT Marathon (see Programme):

http://www.statmt.org/mtm15 and the neural /mtm16

Books:

- Ondřej Bojar: Čeština a strojový překlad. ÚFAL, 2012.
- Philipp Koehn: Statistical Machine Translation. Cambridge University Press, 2009.
  With some slides: http://statmt.org/book/
  NMT: https://arxiv.org/pdf/1709.07809.pdf

Approaches to Machine Translation

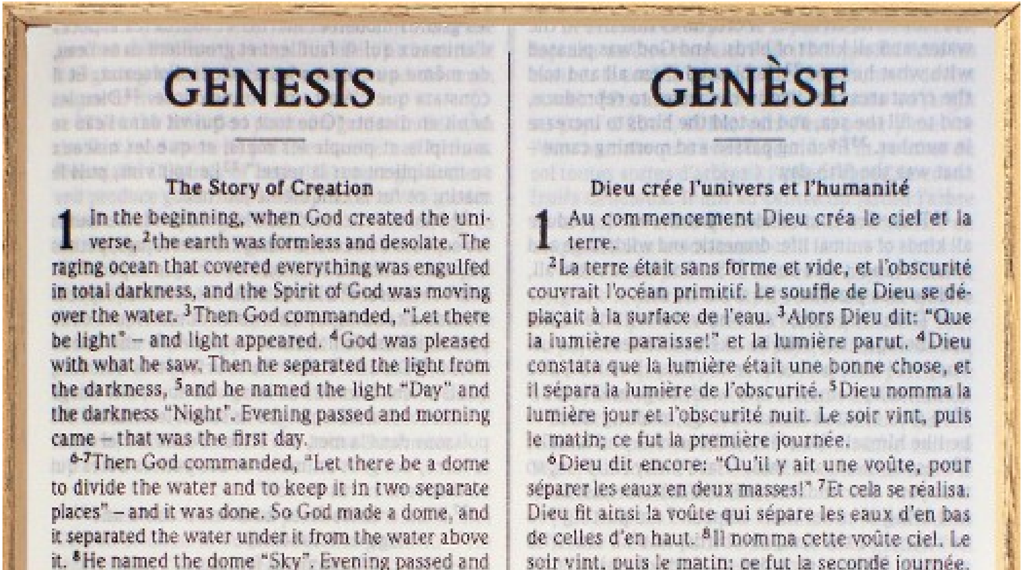


- The deeper analysis, the easier the transfer should be.
- A hypothetical interlingua captures pure meaning.

- Linguistically shallow (direct) methods need just parallel texts.

⇒ They can nicely serve as a tool for transfer.

A Classical Parallel Corpus



GENESIS

The Story of Creation

1 In the beginning, when God created the universe, [2]the earth was formless and desolate. The raging ocean that covered everything was engulfed in total darkness, and the Spirit of God was moving over the water. [3]Then God commanded, "Let there be light" – and light appeared. [4]God was pleased with what he saw. Then he separated the light from the darkness, [5]and he named the light "Day" and the darkness "Night". Evening passed and morning came – that was the first day.

[6-7]Then God commanded, "Let there be a dome to divide the water and to keep it in two separate places" – and it was done. So God made a dome, and it separated the water under it from the water above it. [8]He named the dome "Sky". Evening passed and

GENÈSE

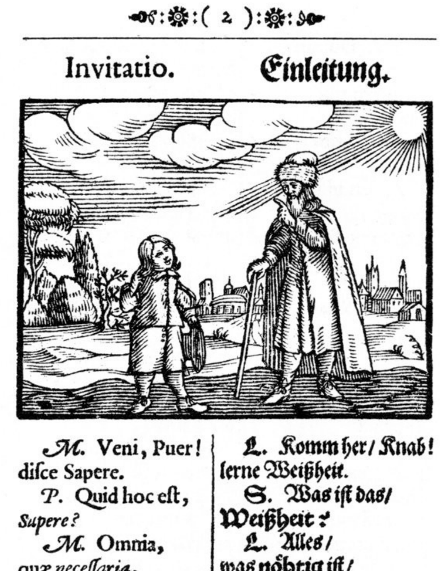
Dieu crée l'univers et l'humanité

1 Au commencement Dieu créa le ciel et la terre.

[2]La terre était sans forme et vide, et l'obscurité couvrait l'océan primitif. Le souffle de Dieu se déplaçait à la surface de l'eau. [3]Alors Dieu dit: "Que la lumière paraisse!" et la lumière parut. [4]Dieu constata que la lumière était une bonne chose, et il sépara la lumière de l'obscurité. [5]Dieu nomma la lumière jour et l'obscurité nuit. Le soir vint, puis le matin; ce fut la première journée.

[6]Dieu dit encore: "Qu'il y ait une voûte, pour séparer les eaux en deux masses!" [7]Et cela se réalisa. Dieu fit ainsi la voûte qui sépare les eaux d'en bas de celles d'en haut. [8]Il nomma cette voûte ciel. Le soir vint, puis le matin; ce fut la seconde journée.

Another Classical One (1658)



(2)

Invitatio.

M. Veni, Puer! diſce Sapere.

P. Quid hoc eſt, *Sapere?*

M. Omnia, *quæ neceſſaria,*

Einleitung.

L. Komm her/ Knab! lerne Weißheit.

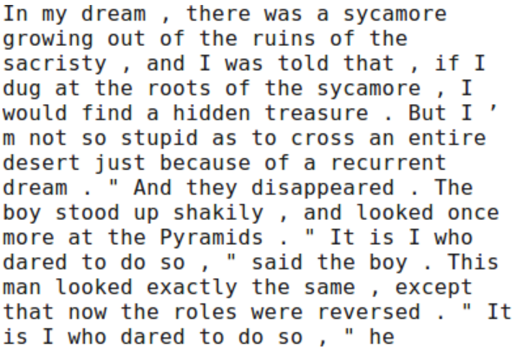
S. Was iſt das/ Weißheit?

L. Alles/ was nöhtig iſt/

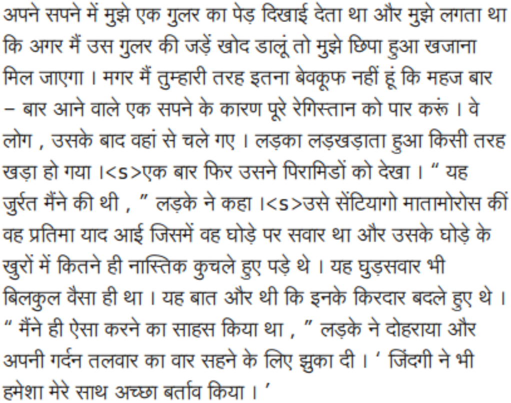
Parallel Corpora

- Web is an immense resource.
- People keep crawling it over and over:
  - Bitextor: Esplà-Gomis and Forcada (2010)
  - http://paracrawl.eu/releases.html (2018)
- Good sources of (multi-)parallel corpora:
  - Corpus OPUS: http://opus.nlpl.eu/
  - WMT tasks data: http://www.statmt.org/wmt19/
  - University-specific corpora, e.g. UFAL released:
    - http://ufal.mff.cuni.cz/czeng (Czech-English)
    - http://ufal.mff.cuni.cz/hindencorp (Hindi-English)
    - http://ufal.mff.cuni.cz/umc/
      (Czech, Russian, Urdu, with English)

From Aligned Documents

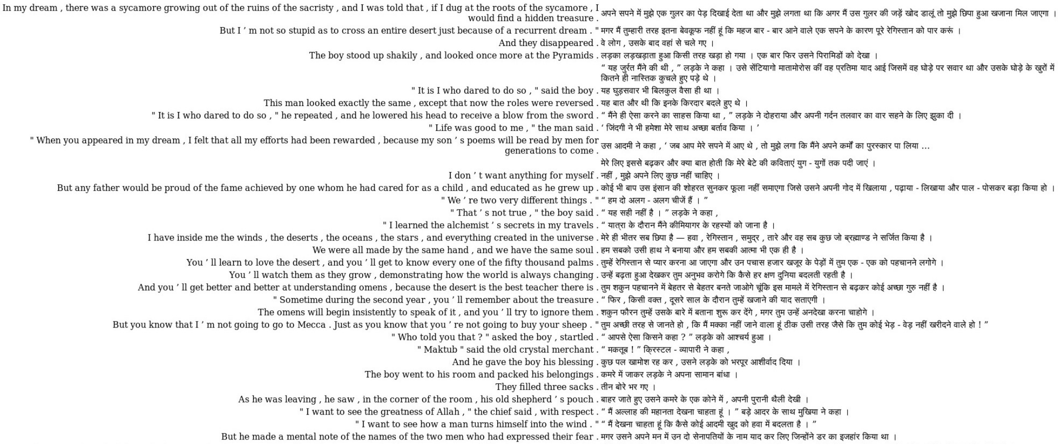


In my dream , there was a sycamore growing out of the ruins of the sacristy , and I was told that , if I dug at the roots of the sycamore , I would find a hidden treasure . But I ’ m not so stupid as to cross an entire desert just because of a recurrent dream . " And they disappeared . The boy stood up shakily , and looked once more at the Pyramids . " It is I who dared to do so , " said the boy . This man looked exactly the same , except that now the roles were reversed . " It is I who dared to do so , " he



अपने सपने में मुझे एक गुलर का पेड़ दिखाई देता था और मुझे लगता था कि अगर मैं उस गुलर की जड़ें खोद डालूं तो मुझे छिपा हुआ खजाना मिल जाएगा । मगर मैं तुम्हारी तरह इतना बेवकूफ नहीं हूं कि महज बार – बार आने वाले एक सपने के कारण पूरे रेगिस्तान को पार करूं । वे लोग , उसके बाद वहां से चले गए । लड़का लड़खड़ाता हुआ किसी तरह खड़ा हो गया ।<s>एक बार फिर उसने पिरामिडों को देखा । “ यह जुर्रत मैंने की थी , ” लड़के ने कहा ।<s>उसे सेंटियागो मातामोरोस कीं वह प्रतिमा याद आई जिसमें वह घोड़े पर सवार था और उसके घोड़े के खुरों में कितने ही नास्तिक कुचले हुए पड़े थे । यह घुड़सवार भी बिलकुल वैसा ही था । यह बात और थी कि इनके किरदार बदले हुए थे । “ मैंने ही ऐसा करने का साहस किया था , ” लड़के ने दोहराया और अपनी गर्दन तलवार का वार सहने के लिए झुका दी । ‘ जिंदगी ने भी हमेशा मेरे साथ अच्छा बर्ताव किया । ’

We Want Sentence Alignment



In my dream , there was a sycamore growing out of the ruins of the sacristy , and I was told that , if I dug at the roots of the sycamore , I would find a hidden treasure .	अपने सपने में मुझे एक गुलर का पेड़ दिखाई देता था और मुझे लगता था कि अगर मैं उस गुलर की जड़ें खोद डालूं तो मुझे छिपा हुआ खजाना मिल जाएगा ।
But I ’ m not so stupid as to cross an entire desert just because of a recurrent dream . "	मगर मैं तुम्हारी तरह इतना बेवकूफ नहीं हूं कि महज बार - बार आने वाले एक सपने के कारण पूरे रेगिस्तान को पार करूं ।
And they disappeared .	वे लोग , उसके बाद वहां से चले गए ।
The boy stood up shakily , and looked once more at the Pyramids .	लड़का लड़खड़ाता हुआ किसी तरह खड़ा हो गया । एक बार फिर उसने पिरामिडों को देखा ।
	“ यह जुर्रत मैंने की थी , ” लड़के ने कहा । उसे सेंटियागो मातामोरोस कीं वह प्रतिमा याद आई जिसमें वह घोड़े पर सवार था और उसके घोड़े के खुरों में कितने ही नास्तिक कुचले हुए पड़े थे ।
" It is I who dared to do so , " said the boy .	यह घुड़सवार भी बिलकुल वैसा ही था ।
This man looked exactly the same , except that now the roles were reversed .	यह बात और थी कि इनके किरदार बदले हुए थे ।
" It is I who dared to do so , " he repeated , and he lowered his head to receive a blow from the sword .	“ मैंने ही ऐसा करने का साहस किया था , ” लड़के ने दोहराया और अपनी गर्दन तलवार का वार सहने के लिए झुका दी ।
" Life was good to me , " the man said .	‘ जिंदगी ने भी हमेशा मेरे साथ अच्छा बर्ताव किया । ’
" When you appeared in my dream , I felt that all my efforts had been rewarded , because my son ’ s poems will be read by men for generations to come .	उस आदमी ने कहा , ‘ जब आप मेरे सपने में आए थे , तो मुझे लगा कि मैंने अपने कर्मों का पुरस्कार पा लिया …
	मेरे लिए इससे बढ़कर और क्या बात होती कि मेरे बेटे की कविताएं युग - युगों तक पदी जाएं ।
I don ’ t want anything for myself .	नहीं , मुझे अपने लिए कुछ नहीं चाहिए ।
But any father would be proud of the fame achieved by one whom he had cared for as a child , and educated as he grew up .	कोई भी बाप उस इंसान की शोहरत सुनकर फूला नहीं समाएगा जिसे उसने अपनी गोद में खिलाया , पढ़ाया - लिखाया और पाल - पोसकर बड़ा किया हो ।
" We ’ re two very different things . "	“ हम दो अलग - अलग चीजें हैं । ”
" That ’ s not true , " the boy said .	“ यह सही नहीं है । ” लड़के ने कहा ,
" I learned the alchemist ’ s secrets in my travels .	“ यात्रा के दौरान मैंने कीमियागर के रहस्यों को जाना है ।
I have inside me the winds , the deserts , the oceans , the stars , and everything created in the universe .	मेरे ही भीतर सब छिपा है — हवा , रेगिस्तान , समुद्र , तारे और वह सब कुछ जो ब्रह्माण्ड ने सर्जित किया है ।
We were all made by the same hand , and we have the same soul .	हम सबको उसी हाथ ने बनाया और हम सबकी आत्मा भी एक ही है ।
You ’ ll learn to love the desert , and you ’ ll get to know every one of the fifty thousand palms .	तुम्हें रेगिस्तान से प्यार करना आ जाएगा और उन पचास हजार खजूर के पेड़ों में तुम एक - एक को पहचानने लगोगे ।
You ’ ll watch them as they grow , demonstrating how the world is always changing .	उन्हें बढ़ता हुआ देखकर तुम अनुभव करोगे कि कैसे हर क्षण दुनिया बदलती रहती है ।
And you ’ ll get better and better at understanding omens , because the desert is the best teacher there is .	तुम शकुन पहचानने में बेहतर से बेहतर बनते जाओगे चूंकि इस मामले में रेगिस्तान से बढ़कर कोई अच्छा गुरु नहीं है ।
" Sometime during the second year , you ’ ll remember about the treasure .	“ फिर , किसी वक्त , दूसरे साल के दौरान तुम्हें खजाने की याद सताएगी ।
The omens will begin insistently to speak of it , and you ’ ll try to ignore them .	शकुन फौरन तुम्हें उसके बारे में बताना शुरू कर देंगे , मगर तुम उन्हें अनदेखा करना चाहोगे ।
But you know that I ’ m not going to go to Mecca . Just as you know that you ’ re not going to buy your sheep .	" तुम अच्छी तरह से जानते हो , कि मैं मक्का नहीं जाने वाला हूं ठीक उसी तरह जैसे कि तुम कोई भेड़ - वेड़ नहीं खरीदने वाले हो ! ”
" Who told you that ? " asked the boy , startled .	“ आपसे ऐसा किसने कहा ? ” लड़के को आश्चर्य हुआ ।
" Maktub " said the old crystal merchant .	“ मकतूब ! ” क्रिस्टल - व्यापारी ने कहा ,
And he gave the boy his blessing .	कुछ पल खामोश रह कर , उसने लड़के को भरपूर आशीर्वाद दिया ।
The boy went to his room and packed his belongings .	कमरे में जाकर लड़के ने अपना सामान बांधा ।
They filled three sacks .	तीन बोरे भर गए ।
As he was leaving , he saw , in the corner of the room , his old shepherd ’ s pouch .	बाहर जाते हुए उसने कमरे के एक कोने में , अपनी पुरानी थैली देखी ।
" I want to see the greatness of Allah , " the chief said , with respect .	“ मैं अल्लाह की महानता देखना चाहता हूं । ” बड़े आदर के साथ मुखिया ने कहा ।
" I want to see how a man turns himself into the wind . "	“ मैं देखना चाहता हूं कि कैसे कोई आदमी खुद को हवा में बदलता है । ”
But he made a mental note of the names of the two men who had expressed their fear .	मगर उसने अपने मन में उन दो सेनापतियों के नाम याद कर लिए जिन्होंने डर का इजहांर किया था ।

Sentence Alignment

Goal: Given a text in two languages, align sentences.
Assume: Sentences hardly ever reordered.

- Classical algorithm: Gale and Church (1993).
  - Based on similar character length of aligned sentences, no words examined.
  - Dynamic-programming search for the best alignment.
  - Allows 0 to 2 sentences in a group: 0-1, 1-0, 1-1, 2-1, 1-2, 2-2.
- Several algorithms for English-Czech evaluated by Rosen (2005).
  - Nearly perfect alignment possible by a combination of aligners.
- The "standard tool": Hunalign (Varga et al., 2005).
- Another option: Gargantua (Braune and Fraser, 2010).

Word Alignment

Goal: Given a sentence in two languages, align words (tokens).
State of the art: GIZA++ (Och and Ney, 2000):

- Unsupervised, only sentence-parallel texts needed.
- Word alignments formally restricted to a function:

$$\text{src token} \mapsto \text{tgt token or NULL}$$

- A cascade of models refining the probability distribution:
  - IBM1: only lexical probabilities: $P(\textit{kočka} = \textit{cat})$
  - IBM3: adds fertility: 1 word generates several others
  - IBM4/HMM: to account for relative reordering
- Only many-to-one links created ⇒ used twice, in both directions.

IBM Model 1

Lexical probabilities:

- Disregard the position of words in sentences.
- Estimated using Expectation-Maximization Loop.

More details available e.g. in MT Marathon slides:

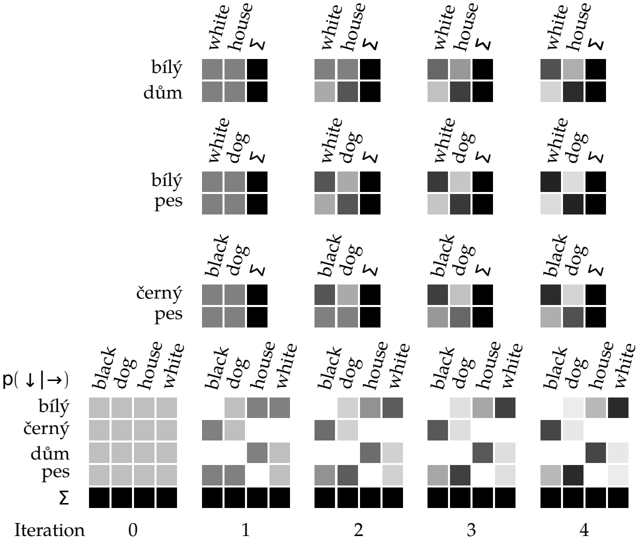
- Aleš Tamchyna.

  `http://www.statmt.org/mtm15`
  → Programme → Tuesday Lecture

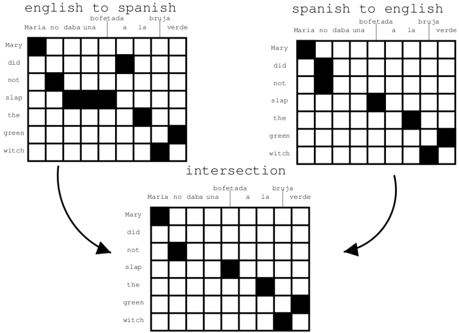
- Patrick Lambert (originally Philipp Koehn)

  `http://lium3.univ-lemans.fr/mtmarathon2010/lectures/02-wordalignment.pdf`

EM Loop in IBM1



Symmetrization for PBMT



“Symmetrization” of two GIZA++ runs:

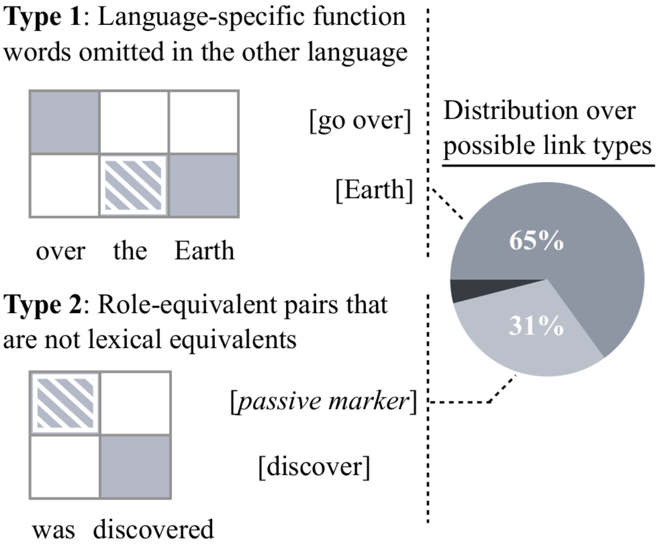
- intersection: high precision, too low recall.
- popular: gdfa (a heuristic between intersection and union).
- minimum-weight edge cover (Matusov et al., 2004).

Troubles with Word Alignment

- Humans have troubles aligning word for word.
    - Mismatch in alignments points 9–18%. (Bojar and Prokopová, 2006)

Top Problematic Words				Top Problematic Parts of Speech			
English		Czech		English		Czech	
361	to	319	,	679	IN	1348	N
259	the	271	se	519	DT	1283	V
159	of	146	v	510	NN	661	R
143	a	112	na	386	PRP	505	P
124	,	74	o	361	TO	448	Z
107	be	61	že	327	VB	398	A
99	it	55	.	310	JJ	280	D
95	that	47	a	245	RB	192	J

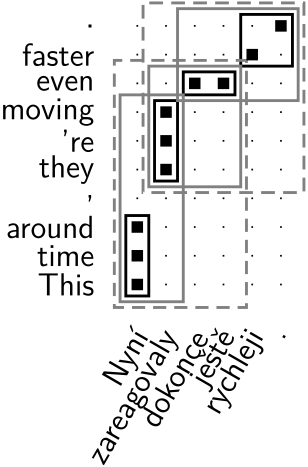
Partial Fix: “Possible” Alignments



Type 1: Language-specific function words omitted in the other language

Type 2: Role-equivalent pairs that are not lexical equivalents

Phrase-Based MT Overview



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychle
...	=	...

Phrase-based MT: choose such segmentation of input string and such phrase “replacements” to make the output sequence “coherent” (3-grams most probable).

1: Align Training Sentences

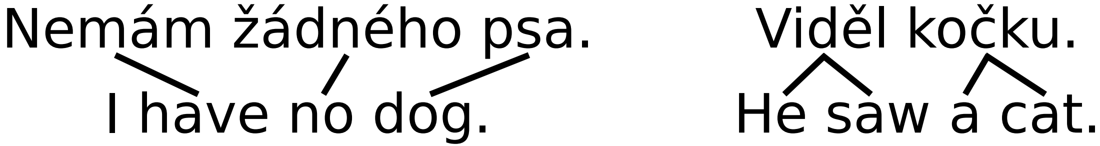
Nemám žádného psa.

I have no dog.

Viděl kočku.

He saw a cat.

2: Align Words



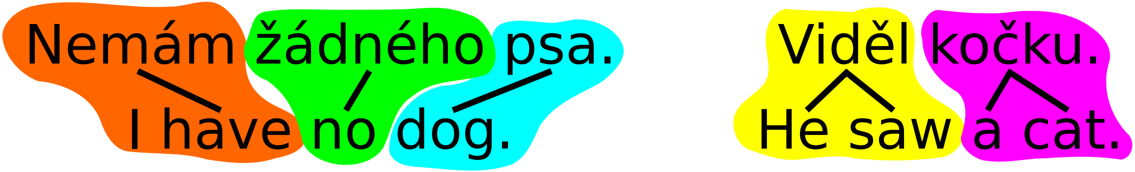
Nemám žádného psa.

I have no dog.

Viděl kočku.

He saw a cat.

3: Extract Phrase Pairs (MTUs)



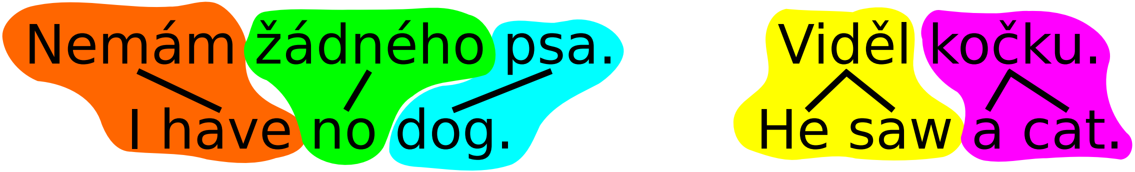
Nemám žádného psa.

I have no dog.

Viděl kočku.

He saw a cat.

4: New Input



Nemám žádného psa.

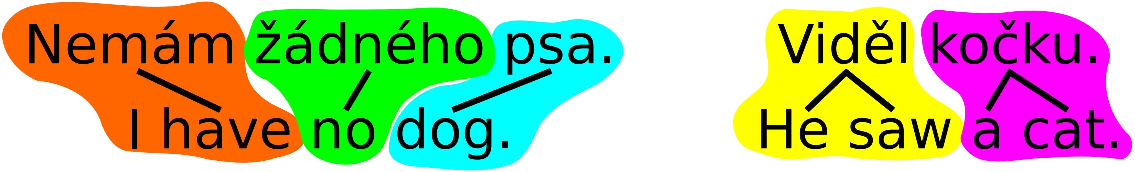
I have no dog.

Viděl kočku.

He saw a cat.

New input: Nemám kočku.

5: Pick Probable Phrase Pairs (TM)



Nemám žádného psa.

I have no dog.

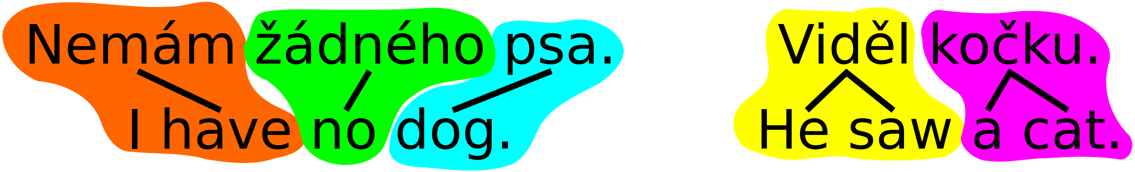
Viděl kočku.

He saw a cat.

... I don't have cat.

New input: Nemám kočku.

6: Cover Input Step-by-Step



Nemám žádného psa.

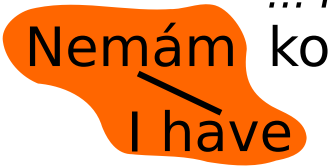
I have no dog.

Viděl kočku.

He saw a cat.

... I don't have cat.

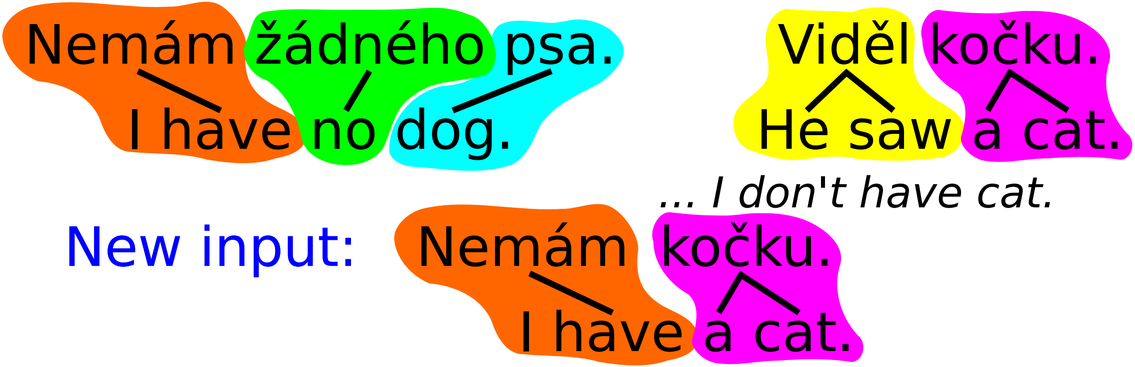
New input: Nemám kočku.



Nemám

I have

7: Read Output



Nemám žádného psa.

I have no dog.

Viděl kočku.

He saw a cat.

New input: Nemám kočku.

I have a cat.

... I don't have cat.

Summary of MT Class 1

- Parallel corpora.
- Sentence alignment.
- Word alignment.
- Phrase-based MT.

Lab: Get GIZA and Moses Running

We will use Eman `http://ufal.mff.cuni.cz/eman`:

- to compile GIZA++ and Moses.
- (optionally) to actually train a PBMT system.

(because GIZA++ needs a patch and Moses needs some flags)
(eman itself may however also need something...)

Bird's Eye View of PBMT

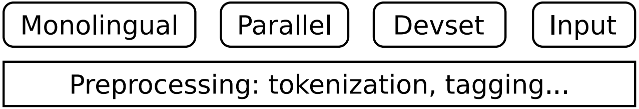
Monolingual

Parallel

Devset

Input

Bird's Eye View of PBMT



Monolingual

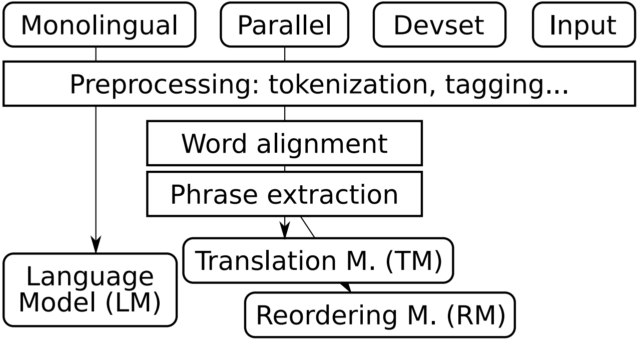
Parallel

Devset

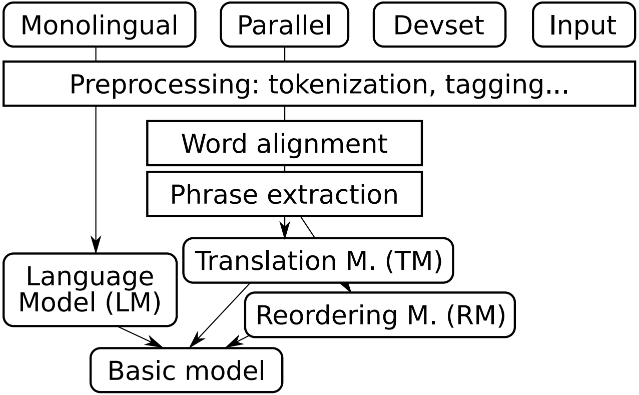
Input

Preprocessing: tokenization, tagging...

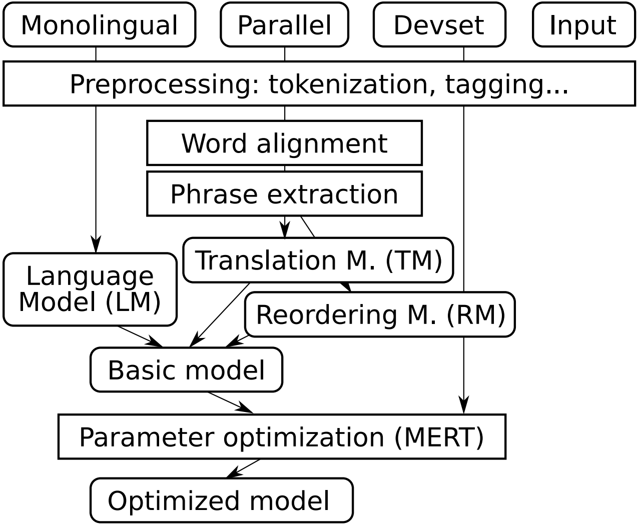
Bird’s Eye View of PBMT



Bird's Eye View of PBMT



Bird’s Eye View of PBMT



Monolingual

Parallel

Devset

Input

Preprocessing: tokenization, tagging...

Word alignment

Phrase extraction

Language Model (LM)

Translation M. (TM)

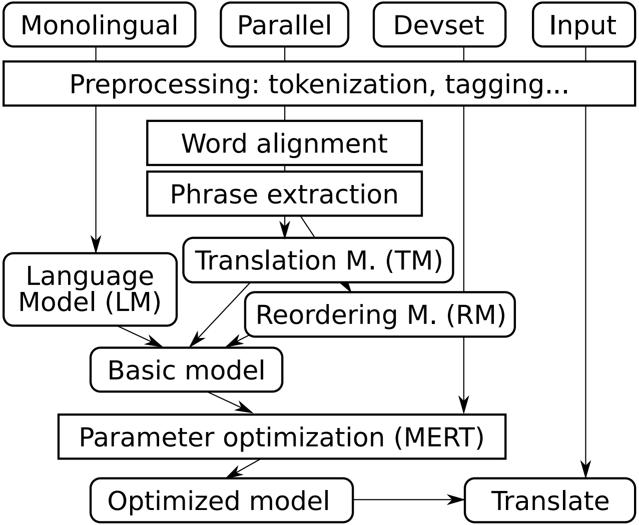
Reordering M. (RM)

Basic model

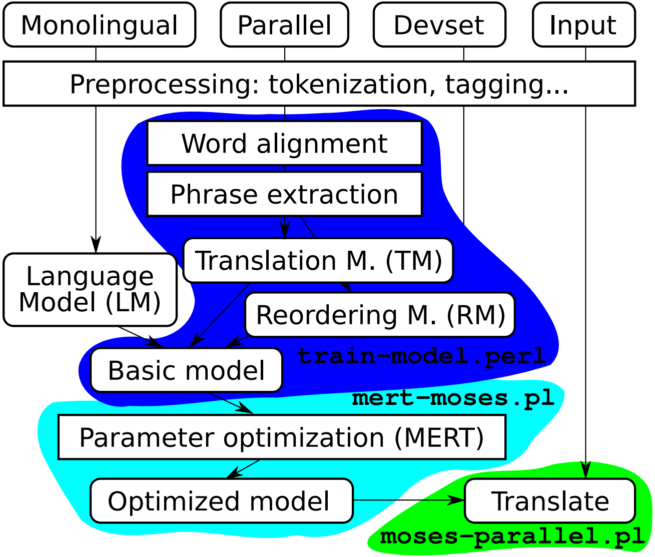
Parameter optimization (MERT)

Optimized model

Bird's Eye View of PBMT



Bird’s Eye View of PBMT



Get Eman, GIZA and Moses

At Student machines, ÚFAL machines, or your laptop:

1. "Install" eman in your home directory:

   ```
   git clone https://redmine.ms.mff.cuni.cz/ufal-smt/eman.git
   ```

2. Make sure eman is in your PATH: Bad things happen if not.

   ```
   export PATH=$HOME/eman/bin/:$PATH
   echo "export PATH=$HOME/eman/bin/:\$PATH" >> ~/.bashrc
   ```

3. Get our SMT Playground (with all the seeds):

   ```
   git clone \
   https://redmine.ms.mff.cuni.cz/ufal-smt/playground.git
   ```

Possibly Fix Perl Dependencies

1. Set up a local Perl repository.

   ```
   http://stackoverflow.com/questions/2980297
   ```

   Copy & paste code from the first answer, just replace `.profile` with `.bashrc`

2. Install the required package:

   ```
   cpanm YAML::XS
   ```

3. Confirm that `eman` runs:

   ```
   eman --man
   ```

Start Compiling Moses

In eman's philosophy, software is just data.

- Binaries should be compiled in timestamped step dirs.
- ...so we know the exact code that was used.

Compile Moses and GIZA++ (all on one line!):

```
BJAMARGS=" link=shared --no-xmlrpc-c
           --max-kenlm-order=12 -a "
  eman init --start mosesgiza
```

✎ Examine the newly created step dir `s.mosesgiza.*`.

Where is the compilation log?
(Moses+GIZA compilation takes ∼8min.)

Use GIZA++

It's easier to run GIZA using my wrapper:

`https://raw.githubusercontent.com/ufal/qtleap/master/cuni_train/bin/gizawrapper.pl`

1. Download gizawrapper, `chmod 755`.
2. Download a parallel corpus of your choice.
   - From OPUS.
   - In “moses” format, which probably means already tokenized.
   - It should have ∼0.1M sentence pairs.
   - Amharic (am) – English (en) Tanzil corpus is a good choice.
3. Run gizawrapper + symal (takes ∼20 min on Am-En):

```
./gizawrapper.pl \
  --bindir=/PATH/TO/playground/s.mosesgiza.????????.2018????-????/bin/ \
  am-en/Tanzil.am-en.am \
  am-en/Tanzil.am-en.?? \
  --dirsym=left,right,int,union | gzip > am-en.ali.gz
```

Observe the Alignment

1. Get my alignment viewer alitextview:

   `http://ufal.mff.cuni.cz/~zabokrtsky/fel/slides/lab09-mt-word-alignment/alitextview.pl`

2. Render the alignment on text console:

```
paste am-en/Tanzil.am-en.am am-en/Tanzil.am-en.en \
      <(zcat am-en.ali.gz ) | cut -f 1,2,5 \
      | ./alitextview.pl | less
```

Improve the Alignment

1. Alignment needs large data.
2. If you don't have it, you need to make statistics denser. Here we lowercase and "stem" (chop words to 4 characters):

```
for f in Tanzil.am-en.??; do
  cat $f | ../playground/scripts/lowercase.pl \
  | ../playground/scripts/stem_factor.pl \
  > $f.lcstem4
done
```

References

Ondřej Bojar and Magdalena Prokopová. 2006. Czech-English Word Alignment. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), pages 1236–1239. ELRA, May.

Fabienne Braune and Alexander Fraser. 2010. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In Coling 2010: Posters, pages 81–89, Beijing, China, August. Coling 2010 Organizing Committee.

John DeNero and Dan Klein. 2010. Discriminative modeling of extraction sets for machine translation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1453–1463, Uppsala, Sweden, July. Association for Computational Linguistics.

Miquel Esplà-Gomis and Mikel L. Forcada. 2010. Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. In Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools, number 93 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(1):75–102.

E. Matusov, R. Zens, and H. Ney. 2004. Symmetric Word Alignments for Statistical Machine Translation. In Proceedings of COLING 2004, pages 219–225, Geneva, Switzerland, August 23–27.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In Proceedings of the 17th conference on Computational linguistics, pages 1086–1090. Association for Computational Linguistics.

Alexandr Rosen. 2005. In Search of Best Method for Sentence Alignment in Parallel Texts. In R. Garabík, editor, Computer Treatment of Slavic and East European Languages, pages 174–185. Veda, Bratislava.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing RANLP 2005