Machine Translation 3: Linguistics in SMT and NMT

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

January 2019
Outline of Lectures on MT

1. Introduction.
   • Why is MT difficult.
   • MT evaluation.
   • Approaches to MT.
   • First peek into phrase-based MT
   • Document, sentence and word alignment.

   • Phrase-based: Assumptions, beam search, key issues.
   • Neural MT: Sequence-to-sequence, attention, self-attentive.

3. Advanced Topics.
   • Linguistic Features in SMT and NMT.
   • Multilinguality, Multi-Task, Learned Representations.
Outline of MT Lecture 3

1. Linguistic features for tokens.
   • Factored phrase-based MT.

2. Linguistic structure to organize search.
   • Non-projectivity.
   • TectoMT: transfer-based deep-syntactic model.

3. Combination to make it actually work.

4. Incorporating linguistic features in NMT.
   • Dedicated models or just data hacks.
     – For multi-task, for multilingual MT.
   • Are the models understanding?
## Morphological Richness (in Czech)

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rich morphology</td>
<td>≥ 4,000 tags possible</td>
<td>50 used</td>
</tr>
<tr>
<td></td>
<td>≥ 2,300 tags seen</td>
<td></td>
</tr>
<tr>
<td>Word order</td>
<td>free</td>
<td>rigid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>News Commentary Corpus</th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>55,676</td>
<td></td>
</tr>
<tr>
<td>Tokens</td>
<td>1.1M</td>
<td>1.2M</td>
</tr>
<tr>
<td>Vocabulary (word forms)</td>
<td>91k</td>
<td>40k</td>
</tr>
<tr>
<td>Vocabulary (lemmas)</td>
<td>34k</td>
<td>28k</td>
</tr>
</tbody>
</table>

Czech tagging and lemmatization: Hajič and Hladká (1998)
English tagging (Ratnaparkhi, 1996) and lemmatization (Minnen et al., 2001).
Morphological Explosion in Czech

MT chooses output words in a form:

- Czech nouns and adjs.: 7 cases, 4 genders, 3 numbers, . . .
- Czech verbs: gender, number, aspect (im/perfective), . . .

<table>
<thead>
<tr>
<th>l</th>
<th>saw</th>
<th>two</th>
<th>green</th>
<th>striped</th>
<th>cats</th>
</tr>
</thead>
<tbody>
<tr>
<td>já</td>
<td>pila</td>
<td>dva</td>
<td>zelený</td>
<td>pruhovaný</td>
<td>kočky</td>
</tr>
<tr>
<td>pily</td>
<td>dvě</td>
<td>zelená</td>
<td>pruhovaná</td>
<td>koček</td>
<td></td>
</tr>
<tr>
<td>. .</td>
<td>dvou</td>
<td>zelené</td>
<td>pruhované</td>
<td>kočkám</td>
<td></td>
</tr>
<tr>
<td>viděl</td>
<td>dvěma</td>
<td>zelení</td>
<td>pruhovaní</td>
<td>kočkách</td>
<td></td>
</tr>
<tr>
<td>viděla</td>
<td>dvěmi</td>
<td>zeleného</td>
<td>pruhovaného</td>
<td>kočkami</td>
<td></td>
</tr>
<tr>
<td>. .</td>
<td></td>
<td>zelených</td>
<td>pruhovaných</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uviděl</td>
<td></td>
<td>zelenému</td>
<td>pruhovanému</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uviděla</td>
<td></td>
<td>zeleným</td>
<td>pruhovaným</td>
<td></td>
<td></td>
</tr>
<tr>
<td>. .</td>
<td></td>
<td>zelenou</td>
<td>pruhovanou</td>
<td></td>
<td></td>
</tr>
<tr>
<td>viděl jsem</td>
<td></td>
<td>zelenými</td>
<td>pruhovanými</td>
<td></td>
<td></td>
</tr>
<tr>
<td>viděla jsem</td>
<td></td>
<td>. .</td>
<td>. .</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Compounding in German:

- Rindfleischetikettierungsüberwachungsaufgabenübertragungs-gesetz.
  “beef labelling supervision duty assignment law”

Agglutination in Hungarian or Finnish:

- istua “to sit down” (istun = “I sit down”)
- istahtaa “to sit down for a while”
- istahdan “I’ll sit down for a while”
- istahtaisin “I would sit down for a while”
- istahtaisinko “should I sit down for a while?”
- istahtaisinkohan “I wonder if I should sit down for a while”
Possible translations differing in morphology:

<table>
<thead>
<tr>
<th>English</th>
<th>Czech 1</th>
<th>Czech 2</th>
<th>Czech 3</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>two green striped cats</td>
<td>dvou zelená pruhovaný kočkách</td>
<td>← garbage</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>dva zelené pruhované kočky</td>
<td>← 3grams ok, 4gram bad</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>dvě zelené pruhované kočky</td>
<td>← correct nominative/accusative</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>dvěma zeleným pruhovaným kočkám</td>
<td>← correct dative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• 3-gram LM too weak to ensure agreement.
• 3-gram LM possibly already too sparse!
Add morphological tag to each output token:

<table>
<thead>
<tr>
<th>two</th>
<th>green</th>
<th>striped</th>
<th>cats</th>
</tr>
</thead>
<tbody>
<tr>
<td>dvou</td>
<td>zelená</td>
<td>pruhovaný</td>
<td>kočkách</td>
</tr>
<tr>
<td>fem-loc</td>
<td>neut-acc</td>
<td>masc-nom-sg</td>
<td>fem-loc</td>
</tr>
<tr>
<td>dva</td>
<td>zelené</td>
<td>pruhované</td>
<td>kočky</td>
</tr>
<tr>
<td>masc-nom</td>
<td>masc-nom</td>
<td>masc-nom</td>
<td></td>
</tr>
<tr>
<td>fem-nom</td>
<td>fem-nom</td>
<td>fem-nom</td>
<td></td>
</tr>
<tr>
<td>dvě</td>
<td>zelené</td>
<td>pruhované</td>
<td>kočky</td>
</tr>
<tr>
<td>fem-nom</td>
<td>fem-nom</td>
<td>fem-nom</td>
<td>fem-nom</td>
</tr>
<tr>
<td>fem-acc</td>
<td>fem-acc</td>
<td>fem-acc</td>
<td>fem-acc</td>
</tr>
<tr>
<td>dvěma</td>
<td>zeleným</td>
<td>pruhovaným</td>
<td>kočkám</td>
</tr>
<tr>
<td>fem-dat</td>
<td>fem-dat</td>
<td>fem-dat</td>
<td>fem-dat</td>
</tr>
</tbody>
</table>
Advantages of Explicit Morphology

- LM over morphological tags generalizes better.
  - \( p(\text{dvě kočkách}) < p(\text{dvě kočky}) \) . . . surely
    But we would need to see all combinations of \( dva \) and \( kočka \)!
    \( \Rightarrow \) Better to ask if \( p(\text{fem-nom fem-loc}) < p(\text{fem-nom fem-nom}) \)
      which is trained on any feminine adj+noun.

- But still does not solve everything.
  - \( p(\text{dvě zelené}) \geq p(\text{dva zelené}) \) . . . bad question anyway!
    Not solved by asking if \( p(\text{fem-nom fem-nom}) \geq p(\text{masc-nom masc-nom}) \).

- Tagset size smaller than vocabulary.
  \( \Rightarrow \) can afford e.g. 7-grams:
  \( p(\text{masc-nom fem-nom fem-nom}) < p(\text{fem-nom fem-nom fem-nom}) \)

Any risks?
Factored Phrase-Based MT

- Both input and output words can have more factors.
- Arbitrary number and order of:

  **Mapping/Translation steps (→)**
  
  Translate (phrases of) source factors to target factors.
  
  two green → dvě zelené

  **Generation steps (↓)**
  
  Generate target factors from target factors.
  
  dvě → *fem-nom*; dva → *masc-nom*
  
  ⇒ Ensures “vertical” coherence.

  **Target-side language models (+LM)**
  
  Applicable to various target-side factors.
  
  ⇒ Ensures “horizontal” coherence.

(Koehn and Hoang, 2007)
Factored Phrase Extraction (1/3)

As in standard phrase-based MT:

1. Run sentence and word alignment,

<table>
<thead>
<tr>
<th>natürlich</th>
<th>hat</th>
<th>naturally</th>
<th>has</th>
<th>fun</th>
<th>with</th>
<th>the</th>
<th>game</th>
</tr>
</thead>
<tbody>
<tr>
<td>hat</td>
<td>naturally</td>
<td>has</td>
<td>fun</td>
<td>with</td>
<td>the</td>
<td>game</td>
<td></td>
</tr>
<tr>
<td>john</td>
<td>naturally</td>
<td>has</td>
<td>fun</td>
<td>with</td>
<td>the</td>
<td>game</td>
<td></td>
</tr>
<tr>
<td>spass</td>
<td>naturally</td>
<td>has</td>
<td>fun</td>
<td>with</td>
<td>the</td>
<td>game</td>
<td></td>
</tr>
<tr>
<td>am</td>
<td>naturally</td>
<td>has</td>
<td>fun</td>
<td>with</td>
<td>the</td>
<td>game</td>
<td></td>
</tr>
<tr>
<td>spiel</td>
<td>naturally</td>
<td>has</td>
<td>fun</td>
<td>with</td>
<td>the</td>
<td>game</td>
<td></td>
</tr>
</tbody>
</table>
Factored Phrase Extraction (2/3)

As in standard phrase-based MT:

1. Run sentence and word alignment,
2. Extract all phrases consistent with word alignment.

\[ \Rightarrow \text{Extracted: natürlich hat john} \rightarrow \text{naturally john has} \]
Factored Phrase Extraction (3/3)

As in standard phrase-based MT:

1. Run sentence and word alignment,
2. Extract same phrases, just another factor from each word.

⇒ Extracted: ADV V NNP → ADV NNP V
Factored Translation Process

Input: (cars, car, NNS)

1. Translation step: lemma ⇒ lemma
   (_, auto, _), (_, automobil, _), (_, vůz, _)

2. Generation step: lemma ⇒ part-of-speech
   (_, auto, N-sg-nom), (_, auto, N-sg-gen), . . . ,
   (_, vůz, N-sg-nom), . . . , (_, vůz, N-sg-gen) . . .

3. Translation step: part-of-speech ⇒ part-of-speech
   (_, auto, N-plur-nom), (_, auto, N-plur-acc), . . . ,
   (_, vůz, N-plur-nom), . . . , (_, vůz, N-sg-gen) . . .

4. Generation step: lemma, part-of-speech ⇒ surface
   (auta, auto, N-plur-nom), (auta, auto, N-plur-acc), . . . ,
   (vozy, vůz, N-plur-nom), . . . , (vozu, vůz, N-sg-gen) . . .
Factored Phrase-Based MT

See slides by Philipp Koehn, pages 49–75:

- Decoding
- Experiments
  - incl. Alternative Decoding Paths
## Translation Scenarios for En→Cs

### Vanilla

<table>
<thead>
<tr>
<th>English</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>form</td>
<td>form</td>
</tr>
<tr>
<td>lemma</td>
<td>lemma</td>
</tr>
<tr>
<td>morphology</td>
<td>morphology</td>
</tr>
</tbody>
</table>

### Translate+Check (T+C)

<table>
<thead>
<tr>
<th>English</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>form</td>
<td>form</td>
</tr>
<tr>
<td>lemma</td>
<td>lemma</td>
</tr>
<tr>
<td>morphology</td>
<td>morphology</td>
</tr>
</tbody>
</table>

### Translate+2·Check (T+C+C)

<table>
<thead>
<tr>
<th>English</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>form</td>
<td>form</td>
</tr>
<tr>
<td>lemma</td>
<td>lemma</td>
</tr>
<tr>
<td>morphology</td>
<td>morphology</td>
</tr>
</tbody>
</table>

### 2·Translate+Generate (T+T+G)

<table>
<thead>
<tr>
<th>English</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>form</td>
<td>form</td>
</tr>
<tr>
<td>lemma</td>
<td>lemma</td>
</tr>
<tr>
<td>morphology</td>
<td>morphology</td>
</tr>
</tbody>
</table>

---

January 2019

MT3: Linguistics in SMT and NMT
Factored Attempts (WMT09)

<table>
<thead>
<tr>
<th>Sents</th>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>Sent/min</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2M</td>
<td>Vanilla</td>
<td>14.24</td>
<td>5.175</td>
<td>12.0</td>
</tr>
<tr>
<td>2.2M</td>
<td>T+C</td>
<td>13.86</td>
<td>5.110</td>
<td>2.6</td>
</tr>
<tr>
<td>84k</td>
<td>T+C+C&amp;T+T+G</td>
<td>10.01</td>
<td>4.360</td>
<td>4.0</td>
</tr>
<tr>
<td>84k</td>
<td>Vanilla MERT</td>
<td>10.52</td>
<td>4.506</td>
<td>–</td>
</tr>
<tr>
<td>84k</td>
<td>Vanilla even weights</td>
<td>08.01</td>
<td>3.911</td>
<td>–</td>
</tr>
</tbody>
</table>

- In WMT07, T+C worked best.
  + fine-tuned tags helped with small data (Bojar, 2007).
- In WMT08, T+C was worth the effort (Bojar and Hajič, 2008).
- In WMT09, our computers could handle 7-grams of forms.
  ⇒ No gain from T+C.
- T+T+G too big to fit and explodes the search space.
  ⇒ Worse than Vanilla trained on the same dataset.
Factored models are “synchronous”, i.e. Moses:
1. Generates fully instantiated “translation options”.
2. Appends translation options to extend “partial hypothesis”.
3. Applies LM to see how well the option fits the previous words.

There are too many possible combinations of lemma+tag.
⇒ Less promising ones must be pruned.
  ! Pruned before the linear context is available.
A Fix: Reverse Self-Training

Goal: Learn from monolingual data to produce new target-side word forms in correct contexts.

<table>
<thead>
<tr>
<th>Source English</th>
<th>Target Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Para 126k</td>
<td>kočka honila. . .</td>
</tr>
<tr>
<td>a cat chased. . .</td>
<td>kočka honit. . . (lem.)</td>
</tr>
<tr>
<td>I saw a cat</td>
<td>viděl jsem kočku</td>
</tr>
<tr>
<td></td>
<td>vidět být kočka (lem.)</td>
</tr>
<tr>
<td>Mono 2M</td>
<td>četl jsem o kočce</td>
</tr>
<tr>
<td>?</td>
<td>číst být o kočka (lem.)</td>
</tr>
</tbody>
</table>

Use reverse translation

I read about a cat ← backed-off by lemmas.

⇒ New phrase learned: “about a cat” = “o kočce”.
The Back-off to Lemmas

- The key distinction from self-training used for domain adaptation (Bertoldi and Federico, 2009; Ueffing et al., 2007).
- We use simply “alternative decoding paths” in Moses:

<table>
<thead>
<tr>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>form</td>
<td>form</td>
</tr>
</tbody>
</table>

  or

<table>
<thead>
<tr>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemma</td>
<td>form</td>
</tr>
</tbody>
</table>

- Other languages (e.g. Turkish, German) need different back-off techniques:
  - Split German compounds.
  - Separate and allow to ignore Turkish morphology.
Small Para, Increasing Mono

BLEU

Monolingual data (mils of sents.)

Mono LM and TM
Mono LM
Increasing Para, Fixed Mono

![Graph showing BLEU scores for different data sets as a function of parallel data size.]

- Mono LM and TM
- Mono LM
- Parallel and Mono
- Parallel

% Test Forms Covered

Parallel data (mils of sents.)

BLEU

January 2019

MT3: Linguistics in SMT and NMT
Summary So Far

- Target-side rich morphology causes data sparseness.
- Factored setups compact the sparseness. 
  but the search space is likely to explode at runtime.
- Explosion contained thanks to pruning.
  but the pruning happens without linear context 
 ⇒ high risk of search errors.

One of possible promising techniques for handling sparseness and avoiding the explosion:

- Reverse self-training (Bojar and Tamchyna, 2011).

... so that was morphology, how about syntax?
Constituency vs. Dependency Trees

Constituency trees (CFG) represent only bracketing: which adjacent constituents are glued tighter to each other.

Dependency trees represent which words depend on which. + usually, some agreement/conditioning happens along the edge.

Constituency

John (loves Mary)
John VP(loves Mary)

Dependency

loves
John Mary

John loves Mary
What Dependency Trees Tell Us

Input: The grass around your house should be cut soon.
Google SMT: Trávu kolem vašeho domu by se měl snížit brzy.
(Google NMT: Tráva kolem vašeho domu by měla být brzy zkrácena.)

• Bad lexical choice for cut = sekat/snížit/krájet/řezat/...
  – Due to long-distance dependency with grass.
  – One can “pump” many words in between.
  – Could be handled by full source-context (e.g. maxent) model.

• Bad case of tráva.
  – Depends on the chosen active/passive form:
    active⇒accusative    passive⇒nominative
    trávu . . . byste se měl posekat  tráva . . . by se měla posekat
    tráva . . . by měla být posekána

Examples by Zdeněk Žabokrtský, Karel Oliva and others.
• Tree context (neighbours in the dependency tree):
  – is better at predicting lexical choice than $n$-grams.
  – often equals linear context:
    - Czech manual trees: 50% of edges link neighbours,
      80% of edges fit in a 4-gram.

• Phrase-based MT is a very good approximation.

• Hierarchical MT (phrases with gaps) can even capture the dependency in one phrase:
  \[ X \rightarrow < \text{the grass } X \text{ should be cut, trávu } X \text{ byste měl posekat} > \]
“Crossing Brackets”

- Constituent outside its father’s span causes “crossing brackets.”
  - Linguists use “traces” (□) to represent this.
- Sometimes, this is not visible in the dependency tree:
  - There is no “history of bracketing”.
  - See Holan et al. (1998) for dependency trees including derivation history.

Despite this shortcoming, CFGs are popular and “the” formal grammar for many. Possibly due to the charm of the father of linguistics, or due to the abundance of dependency formalisms with no clear winner (Nivre, 2005).
Non-Projectivity

= a gap in a subtree span, filled by a node higher in the tree.

Ex. Dutch “cross-serial” dependencies, a non-projective tree with one gap caused by saw within the span of swim.

\[
\begin{array}{ccccccc}
\ldots \text{dat} & \text{Jan} & \underline{\text{kinderen}} & \underline{\text{zag}} & \underline{\text{zwemmen}} \\
\ldots \text{that} & \text{John} & \underline{\text{children}} & \underline{\text{saw}} & \underline{\text{swim}} \\
\ldots \text{that John saw children swim.}
\end{array}
\]

- 0 gaps \(\Rightarrow\) projective tree \(\Rightarrow\) can be represented in a CFG.
- \(\leq 1\) gap & “well-nested” \(\Rightarrow\) mildly context sensitive (TAG).

See Kuhlmann and Möhl (2007) and Holan et al. (1998).
Why Non-Projectivity Matters?

• CFGs cannot handle non-projective constructions:

Imagine John **grass** saw **being-cut**!

• No way to glue these crossing dependencies together:
  
  – Lexical choice:

    \[ X \rightarrow< \text{grass } X \text{ being-cut, trávu } X \text{ sekat } > \]
  
  – Agreement in gender:

    \[ X \rightarrow< \text{John } X \text{ saw, Jan } X \text{ viděl } > \]
    \[ X \rightarrow< \text{Mary } X \text{ saw, Marie } X \text{ viděla } > \]

• Phrasal chunks can memorize fixed sequences containing:
  
  – the non-projective construction
  
  – and all the words in between! (⇒ extreme sparseness)
Is Non-Projectivity Severe?

Depends on the language.

In principle:

- Czech allows long gaps as well as many gaps in a subtree.

In treebank data:

- 23% of Czech sentences contain a non-projectivity.
- 99.5% of Czech sentences are well nested with \( \leq 1 \) gap.
Tectogrammatics: Deep Syntax Culminating

Background: Prague Linguistic Circle (since 1926).

Materialized theory — Treebanks:
- Czech: PDT 1.0 (2001), PDT 2.0 (2006)
- Czech-English: PCEDT 1.0 (2004), PCEDT 2.0 (2012)

Practice — Tools:
- parsing Czech to surface: McDonald et al. (2005)
- parsing Czech to deep: Klimeš (2006)
- parsing English to surface: well studied (+rules convert to dependency trees)
- parsing English to deep: heuristic rules (manual annotation in progress)
- generating Czech surface from t-layer: Ptáček and Žabokrtský (2006)
Layers in PDT
Analytical vs. Tectogrammatical

- hide auxiliary words, add nodes for “deleted” participants
- resolve e.g. active/passive voice, analytical verbs etc.
- “full” t-layer resolves much more, e.g. topic-focus articulation or anaphora
To by se mělo změnit.

#45 This should be changed.
Czech and English T-Layer

Predicate-argument structure: change_{should}(ACT: someone, PAT: it)

January 2019   MT3: Linguistics in SMT and NMT
The Tectogrammatical Hope

Transfer at t-layer should be easier than direct translation:

- Reduced structure size (auxiliary words disappear).
- Long-distance dependencies (non-projectivites) solved at t-layer.
- Word order ignored / interpreted as information structure (given/new).
- Reduced vocabulary size (Czech morphological complexity).
- Czech and English t-trees structurally more similar ⇒ less parallel data might be sufficient (but more monolingual).
- Ready for fancy t-layer features: co-reference.

The complications:

- 47 pages documenting data format (PML, XML-based, sort of typed)
- 1200 pages documenting Czech t-structures "Not necessary" once you have a t-tree but useful understand or to blame the right people.
“TectoMT Transfer” (1/3)

ANALYSIS

deep syntax: tectogrammatical layer

shallow syntax: analytical layer

morphological layer

source language (English)

TRANSFER

t-layer

SYNTHESIS

a-layer

m-layer

w-layer

target language (Czech)
“TectoMT Transfer” (2/3)

**ANALYSIS**
- tectogrammatical layer
  - fill formems
  - grammatemes
  - build t-tree
  - mark edges to contract

**TRANSFER**
- query dictionary
- use HMTM

**SYNTHESIS**
- t-layer
  - fill morphological categories
  - impose agreement
  - add functional words

- a-layer
  - generate wordforms

- m-layer
  - concatenate

- w-layer

**morphological layer**
- parser (McDonald's MST)
- tagger (Morce)
- lemmatization
- tokenization
- segmentation

**analytical layer**
- analytical functions

**rule based & statistical blocks**

source language (English)
target language (Czech)
“TectoMT Transfer” (3/3)

To learn more: Slides 6–28 by Martin Popel (2009):

- Illustration of TectoMT transfer.
- Analysis of translation errors.
- Hidden Markov Tree Model (HMTM).

**Bad news:** TectoMT alone performs poorly.

- Errors cummulate.
- T-layer does bring its independence assumptions.
- No means for plain copy-paste.
Poor Man’s System Combination

- Translate input with TectoMT.
- Align translation back to source.
- Extract phrases.
- Add as a separate phrase table.
- MERT to find weights of both phrase tables.
TectoMT Brings Phrases

Input  I saw two green striped cats.

TectoMT Output  Viděl jsem dvě zelené pruhované kočky.

Phrases extracted:

I saw  =  Viděl jsem
I saw two  =  Viděl jsem dvě
...  =  ...
two  =  dvě
two green  =  dvě zelené
two green striped  =  dvě zelené pruhované
two green striped cats  =  dvě zelené pruhované kočky
...  =  ...
TectoMT Brings Phrases

The output of TectoMT covers (most of) the source.

• Long and short phrases, one form only.

<table>
<thead>
<tr>
<th>I saw</th>
<th>two</th>
<th>green</th>
<th>striped</th>
<th>cats</th>
</tr>
</thead>
<tbody>
<tr>
<td>já</td>
<td>pila</td>
<td>dva</td>
<td>zelený</td>
<td>pruhovaný kočky</td>
</tr>
<tr>
<td>pily</td>
<td>dvě</td>
<td>zelená</td>
<td>pruhovaná koček</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>dvou</td>
<td>zelené</td>
<td>pruhované kočkám</td>
<td></td>
</tr>
<tr>
<td>viděl</td>
<td>dvěma</td>
<td>zelení</td>
<td>pruhovaní kočkách</td>
<td></td>
</tr>
<tr>
<td>viděla</td>
<td>dvěmi</td>
<td>zeleného</td>
<td>pruhovaného kočkami</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>zelených</td>
<td>pruhovaných</td>
<td></td>
</tr>
<tr>
<td>viděl jsem</td>
<td></td>
<td>zelenými</td>
<td>pruhovanými</td>
<td></td>
</tr>
<tr>
<td>viděla jsem</td>
<td></td>
<td></td>
<td></td>
<td>. . .</td>
</tr>
</tbody>
</table>
TectoMT Brings Phrases

The output of TectoMT covers (most of) the source.

- Long and short phrases, one form only.

<table>
<thead>
<tr>
<th>l</th>
<th>saw</th>
<th>two</th>
<th>green</th>
<th>striped</th>
<th>cats</th>
</tr>
</thead>
</table>
| já      | pila| dva | zelený| pruhovaný| kočky|.
| pily    | dve | zelená| pruhovaná| kočky|
| . . .   | dve | zelené| pruhované | koček|
| viděl   | dvou| zelené| pruhované | kočkám|
| viděla  | dvěma| zelení| pruhovaní | kočkách|
| . . .   | dvěmi| zeleného| pruhovaného | kočkami|
| viděl jsem |  |  | zelených | pruhovaných |
| viděl jsem |  | dvě | zelené | pruhované kočky |
| viděla jsem |  |  | dvě zelené | pruhované kočky |
Chimera: Complex Combination

Chimera (자동차+철도차량+어선) was beating everyone in 2013–2015.

• Input:
  – Famous cases also relate to graphic elements.
• TectoMT translates using deep syntax:
  – Slavné případy se být týkají grafické prvky.
• PBMT adds 200M en-cs sents and 3,6G cs words:
  – Slavné případy se týkají také grafické prvky.
• Automatic error correction for agreement or negation:
  – Slavné případy se týkají také grafických prvků.
• Google SMT: Slavné případy týkat i grafické prvky.
• Google NMT: Slavné případy se také týkají grafických prvků.
Summary So Far

• Meaning of sentences is usually *compositional*.
• Syntax describes the composition.
  – Expressed with various surface features (e.g. case).
  – Syntactic context more important than linear context.
  – Non-projectivity: composition $\neq$ concatenation.
• Syntax comes at a cost:
  – Theory you have to learn.
  – More complex search space.
  – Cummulation of errors.
• Syntactic SMT did not outperform PBMT in general.
  – We successfully utilized syntax only within PBMT.
And Now for Something...

Remember: 
\[ p(e_1^I | f_1^J) = p(e_1 | f_1^J) \cdot p(e_2 | e_1, f_1^J) \cdot p(e_3 | e_2, e_1, f_1^J) \cdots \]
Some Advanced Topics in NMT

• Self-Attention
• Linguistic Features in NMT.
• Multi-Task Training.
• Multi-Lingual MT.

These can be done with:

a) dedicated architectures, e.g. Eriguchi et al. (2017)
b) **hacked input/output for seq2seq.**

• Learned Representations.
Self-Attention (Transformer Model)


Three uses of multi-head attention in Transformer

- **Encoder-Decoder Attention:**
  - $Q$: previous decoder layers; $K = V$: outputs of encoder
  - Decoder positions attend to all positions of the input.

- **Encoder Self-Attention:**
  - $Q = K = V$: outputs of the previous layer of the encoder
  - Encoder positions attend to all positions of previous layer.

- **Decoder Self-Attention:**
  - $Q = K = V$: outputs of the previous decoder layer.
  - Masking used to prevent depending on future outputs.
  - Decoder attends to all its previous outputs.
Linguistic Features in NMT

• Source word factors easy to incorporate:
  – Concatenate embeddings of the various factors.
  – POS tags, morph. features, source dependency labels help en→de and en→ro (Sennrich and Haddow, 2016).

• Target word factors:
  – Interleave for morphology: (Tamchyna et al., 2017)
    
    | Source | Target |
    |--------|--------|
    | there are a million different kinds of pizza. | existují miliony druhů piz@zy. |
    | Baseline (BPE) | Interleave |
    | existovat NNIP1 milion NNIP2 druh NNFS2 pizza Z: |

  – Interleave for syntax: (Nadejde et al., 2017)

    | Source | Target |
    |--------|--------|
    | Obama receives Net+ an+ yahu in the capital of USA | NP Obama ((S[ocl]\NP)/PP)/NP receives NP Net+ an+ yahu PP/NP in |
Suspicious Results on Multi-Tasking

My students Dan Kondratyuk and Ronald Cardenas retried Nadejde et al. (2017) with:

- sequence-to-sequence model,
- Transformer model.

Predicting target syntax using:

- a secondary decoder
  (The sequence of CCG tags may not match the translated sentence.)
- interleaving.

As tags, they used:

- correct CCG tags,
- random tags,
- a single dummy tag.
Suspicious Results on Multi-Tasking

![Graph showing training steps (millions) vs. Seq2seq performance for different methods: Baseline, CCG, Random, and Same. The graph indicates that the CCG method performs similar to the Baseline, while the Random and Same methods show different performance trends.]

January 2019

MT3: Linguistics in SMT and NMT
Suspicious Results on Multi-Tasking

![Diagram showing the performance of different models over training steps. The x-axis represents training steps (millions), and the y-axis represents Multi-Decoder. The line graph compares Baseline, CCG, Random, and Same models.]

January 2019
MT3: Linguistics in SMT and NMT
Suspicious Results on Multi-Tasking

![Graph showing training steps (millions) vs. interleaved performance for different methods: Baseline, CCG, Random, Same. The graph indicates that Random performs similarly to the baseline, while CCG shows a deviation from the baseline.]

January 2019
MT3: Linguistics in SMT and NMT
Suspicious Results on Multi-Tasking

![Graph showing training steps vs. Multi-Decoder performance](image)

- Baseline
- CCG
- Random
- Same

January 2019

MT3: Linguistics in SMT and NMT

53
Multi-Lingual MT

... simply feed in various language pairs.

<table>
<thead>
<tr>
<th>Source Sent 1 (De)</th>
<th>2en versetzen Sie sich mal in meine Lage!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Sent 1 (En)</td>
<td>put yourselves in my position</td>
</tr>
<tr>
<td>Source Sent 2 (En)</td>
<td>2nl I flew on Air Force Two for eight years</td>
</tr>
<tr>
<td>Target Sent 2 (Nl)</td>
<td>ik heb acht jaar lang met de Air Force Two gevlogen</td>
</tr>
</tbody>
</table>

• The model of the same size will learn both pairs.
• Hopefully benefiting from various similarities.
• Risk of catastrophic forgetting.

See Johnson et al. (2016) or Ha et al. (2017).
Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When “simpler” mean “shorter”:
  - Clear jumps in score as bins of longer sentences are allowed.
  - Reversed curriculum unlearns to produce long sentences.

January 2019
MT3: Linguistics in SMT and NMT
## Surprising Results with Multiling.

### Transfer

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Baseline BLEU</th>
<th>Baseline Steps</th>
<th>Direct transfer BLEU</th>
<th>Direct transfer Steps</th>
<th>Transformed vocab BLEU</th>
<th>Transformed vocab Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-to-Odia</td>
<td>3.54</td>
<td>45k</td>
<td>0.04</td>
<td>47k</td>
<td>6.38</td>
<td>38k</td>
</tr>
<tr>
<td>English-to-Estonian</td>
<td>8.13</td>
<td>95k</td>
<td>14.48</td>
<td>180k</td>
<td>14.18</td>
<td>175k</td>
</tr>
<tr>
<td>English-to-Finnish</td>
<td>14.42</td>
<td>420k</td>
<td>16.12</td>
<td>255k</td>
<td>16.73</td>
<td>270k</td>
</tr>
<tr>
<td>English-to-German</td>
<td>36.72</td>
<td>270k</td>
<td>38.58</td>
<td>190k</td>
<td>39.28</td>
<td>110k</td>
</tr>
<tr>
<td>English-to-Russian</td>
<td>27.81</td>
<td>1090k</td>
<td>25.50</td>
<td>630k</td>
<td>28.65</td>
<td>450k</td>
</tr>
<tr>
<td>English-to-French</td>
<td>33.72</td>
<td>820k</td>
<td>34.41</td>
<td>660k</td>
<td>34.46</td>
<td>720k</td>
</tr>
<tr>
<td>French-to-Spanish</td>
<td>31.10</td>
<td>390k</td>
<td>31.55</td>
<td>435k</td>
<td>31.67</td>
<td>375k</td>
</tr>
</tbody>
</table>

Best score and lowest training time in each row in bold.

- Reusing the knowledge of English source can help really a lot.
- Pre-training Transformer on fully unrelated language pair can help, too.
Learned Representations

- Deep learning researchers easily claim that NNs learn the meaning of the sentences.
- This is possible, but not achieved in practice, yet:

![Translation Interface]

- Máma mele maso?
- Máma maso mele?
- Mele máma maso?
- Mele maso máma?
- Maso mele máma?
- Maso máma mele?

- Mum is mincing meat?
- Mommy meat?
- My mom's meat?
- My Flesh Mum?
- My mom's meat?
legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista svetokorbel, který odpovídal na otázky novinek.

**Human Output:**
slovenská punková legenda extip se vrátila
legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. Soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. Své zkušenosti s tím má kytarista svetokorbel, který odpovídal na otázky novinek.

**Human Output:**
slovenská punková legenda extip se vrátila

“Summarized” by Google Transformer Model:
slovenská kapela extip se vrací do prahy
Meaning Understood?

Input:
legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista svetokorbel, který odpovídal na otázky novinek.

Human Output:
slovenská punková legenda extip se vrátila

“Summarized” by Google Transformer Model:
slovenská kapela extip se vrací do prahy
Meaning Understood? Surely Not.

na strahově slovenská kapela extip se vrací do prahy
v o2 aréně slovenská kapela extip se vrací do prahy
na hradecku slovenská kapela extip se vrací do čech
u vajgaru slovenská kapela extip se vrací do prahy
Not Understood.

na strahově slovenská kapela extip se vrací do prahy
v o2 aréně slovenská kapela extip se vrací do prahy
na hradecku slovenská kapela extip se vrací do čech
u vajgaru slovenská kapela extip se vrací do prahy
ve stromovce slovenská kapela extip se vrací na scénu.
tentokrát kvůli drogám v reedici. s. s. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. i. m. . . . . . . . . . . . . . . m. . . m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m.

January 2019 MT3: Linguistics in SMT and NMT
Many More Details

... see the 234 slides (ACL 2016 tutorial, 58MB):

https://sites.google.com/site/acl16nmt/

The basics of NMT are here:

• slides 14–19, 24–25: NMT for one word, overview.
• slides 47–53: Recurrent neural LM.
• slides 84–95: Encoder-decoder, decoding.
• slides 130–140: Encoder-decoder with attention.
• slides 192–204: Multi-task and multi-lingual.
• ... but also the basics of NN, e.g. GRU (slides 72–79).
Summary

Linguistic features added:

- as factors (word-level annotations) to phrase-based MT
- as deep syntax, organizing the whole process.
- as source factors to NMT.
- as secondary tasks to NMT.

SMT (and transfer-based MT) suffer from unjustified assumptions.

Neural networks:

- get rid of most of the assumptions.
- but are very expensive to train.
- and it is still not clear how much generalization is learned.
References


References

Dependency-Based Grammars, Montreal. University of Montreal.


References


