Alignment

Ondřej Bojar

March 19, 2020
Outline

- CzEng (http://ufal.mff.cuni.cz/czeng)
  - Sources of (Czech-English) Parallel Texts.
  - Licensing Issues.
  - Impact of Data Type on MT Quality Gain.
- Mining the Web.
- Document Alignment.
- Sentence Alignment.
- Word Alignment.
  - IBM Model 1 and the Expectation-Maximization Loop.
- Problems of Word Alignment.
- Tectogrammatical Alignment.
Overview of Phrase-Based MT

1. Given parallel word-aligned corpus,
2. Extract phrases consistent with word alignment,
3. Translate by replacing phrases.

...but how to do 1?
Data Acquisition
Sources of Texts in CzEng 0.7

Legal texts:
- Acquis Communautaire Parallel Corpus
- The European Constitution proposal from the OPUS corpus
- samples from the Official Journal of the European Union

Stories and Commentaries:
- Readers’ Digest stories
- e-books: Project Gutenberg and Palmknihy.cz and a subset of the Kačenka parallel corpus
- articles from Project Syndicate

User-supplied data: ...not always complete sentences
- Czech localization of KDE and GNOME open-source projects
- user-contributed translations from the Navajo project
## Texts in CzEng 0.7 – Data Sizes

<table>
<thead>
<tr>
<th>Text Type</th>
<th>Sentences</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquis Communautaire</td>
<td>64.1%</td>
<td>69.0%</td>
</tr>
<tr>
<td>Readers’ Digest</td>
<td>8.6%</td>
<td>8.6%</td>
</tr>
<tr>
<td>Project Syndicate</td>
<td>6.5%</td>
<td>8.9%</td>
</tr>
<tr>
<td><strong>KDE Messages</strong></td>
<td><strong>6.2%</strong></td>
<td><strong>1.9%</strong></td>
</tr>
<tr>
<td><strong>GNOME Messages</strong></td>
<td><strong>5.7%</strong></td>
<td><strong>1.9%</strong></td>
</tr>
<tr>
<td>Kačenka</td>
<td>4.2%</td>
<td>4.9%</td>
</tr>
<tr>
<td><strong>Navajo User Translations</strong></td>
<td><strong>2.3%</strong></td>
<td><strong>2.1%</strong></td>
</tr>
<tr>
<td>E-Books</td>
<td>1.2%</td>
<td>1.6%</td>
</tr>
<tr>
<td>European Constitution</td>
<td>0.8%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Samples from European Journal</td>
<td>0.4%</td>
<td>0.5%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1.4 mil.</strong></td>
<td><strong>21 mil.</strong></td>
</tr>
</tbody>
</table>

Community-supplied data in bold.
The Navajo Project

- Anonymous contributors correct MT output of Wikipedia texts.
- About 2,000 segments used to be generated each month.
- Manual evaluation of 1,000 randomly selected segments:

<table>
<thead>
<tr>
<th>Translation Quality</th>
<th>Proportion in the Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>precise, flawless</td>
<td>69.0%</td>
</tr>
<tr>
<td>not translated</td>
<td>6.8%</td>
</tr>
<tr>
<td>incomplete</td>
<td>6.6%</td>
</tr>
<tr>
<td>imprecise</td>
<td>5.8%</td>
</tr>
<tr>
<td>precise, almost flawless</td>
<td>4.5%</td>
</tr>
<tr>
<td>machine-generated</td>
<td>4.4%</td>
</tr>
<tr>
<td>vandalism</td>
<td>2.7%</td>
</tr>
<tr>
<td>other</td>
<td>0.2%</td>
</tr>
</tbody>
</table>
KDE and GNOME Localizations

- Two major open-source software projects,
- Contributors not anonymous $\Rightarrow$ the quality considerably higher (almost professional)
- Only rarely full sentences, mostly short system messages and user interface elements e.g. “OK”, “Yes” or “Delete file”
Licensing Issues

- Much more data are available on the Internet,
- Only a fraction labelled for reuse.

<table>
<thead>
<tr>
<th>Source of Texts and Translation</th>
<th>cs</th>
<th>en</th>
<th>cs</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community Transl. of Proprietary Texts</td>
<td>19.5M</td>
<td>25.3M</td>
<td>37.8%</td>
<td>41.1%</td>
</tr>
<tr>
<td>Professional</td>
<td>21.3M</td>
<td>23.9M</td>
<td>41.2%</td>
<td>38.9%</td>
</tr>
<tr>
<td>Proprietary</td>
<td>9.6M</td>
<td>10.9M</td>
<td>18.6%</td>
<td>17.7%</td>
</tr>
<tr>
<td>Community</td>
<td>1.2M</td>
<td>1.4M</td>
<td>2.4%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Total</td>
<td>51.6M</td>
<td>61.5M</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

CzEng 0.7 ≈ Professional + Community sources; in bold
Training Data Composition

In-domain Professional Translation

Community-supplied with proper copyright

Out-of-domain Professional Translation

Community-supplied, copyright unclear

In-domain Professional Translation

Community-supplied with proper copyright

Out-of-domain Professional Translation

Community-supplied, copyright unclear
OOV and PBMT Quality In/Out of Domain
Community Data Out-of-Domain

![Graph showing OOV Rate vs BLEU with categories: domain-D, out-of-domain-D, comm, no copy: X, Proprietary]
Professional Out-of-Domain

![Graph showing OOV Rate vs BLEU with different categories: domain-D, out-of-domain-D, comm, no copy: X, CC, Proprietary.]

- **OOV Rate**: Measured along the y-axis.
- **BLEU**: Measured along the x-axis.
- **Categories**:
  - Domain-D
  - Out-of-domain-D
  - Communication, no copy: X
  - CC
  - Proprietary

Legend:
- Green: Domain-D
- Blue: Out-of-domain-D
- Red: Communication, no copy: X
- Black: CC
- Gray: Proprietary
Everything Out-of-Domain

![Diagram showing OOV Rate vs BLEU for different categories: domain-D, out-of-domain-D, comm, no copy: X, Proprietary. The categories are represented by different colors and sizes, indicating varying performance metrics.]
Similar Volume of in-Domain: Much Better
Additional Data Improve Coverage
But Out-of-Domain Can Decrease Quality
Applying Out of Domain? Much Worse.
More Data → Better Coverage

![Graph showing the relationship between OOV Rate and BLEU scores for different categories: domain-D, out-of-domain-D, and proprietary, no copy: X. The graph includes data points labeled with C, D, CX, CPX, and DP, indicating different conditions or metrics.](image-url)
...But Not Much Better Quality

![Graph showing OOV Rate vs. BLEU for different scenarios and domains.](image)

- OOV Rate
- BLEU
- Proprietary, comm, no copy: X
- in D, prof.
- CC
- domain-D
- out-of-domain-D
- C
- CX
- P
- CPX
- D
- DPCX
- DCPX
CzEng Releases 2006–2020

- Reached 180M million sentence pairs:
  - 0.6 cs / 0.7 en gigawords of genuine parallel text (61M sentpairs)
  - 2.0 cs / 2.3 en gigawords of synthetic text (127M sentpairs)

<table>
<thead>
<tr>
<th>Ver.</th>
<th>S. Pairs</th>
<th>Main Focus</th>
<th>Details in</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.9M</td>
<td>Sentence alignment, common format</td>
<td>Bojar and Žabokrtský (2006)</td>
</tr>
<tr>
<td>0.7</td>
<td>1.0M</td>
<td>Used in WMT06 and WMT07</td>
<td>Bojar et al. (2008)</td>
</tr>
<tr>
<td>0.9</td>
<td>8.0M</td>
<td>Automatic annotation up to t-layer</td>
<td>Bojar and Žabokrtský (2009)</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>Sentence-level filtering</td>
<td>Bojar et al. (2010)</td>
</tr>
<tr>
<td>1.0</td>
<td>15.0M</td>
<td>Improving monolingual annotation through parallel data</td>
<td>Bojar et al. (2012)</td>
</tr>
<tr>
<td>1.6</td>
<td>62.5M</td>
<td>Processing tools dockered</td>
<td>Bojar et al. (2016)</td>
</tr>
<tr>
<td>1.7</td>
<td>57.1M</td>
<td>Block-level filtering</td>
<td>–</td>
</tr>
<tr>
<td>2.0</td>
<td>188.0M</td>
<td>Filtering + Synthetic data</td>
<td>–</td>
</tr>
</tbody>
</table>
Methods
Mining the Web

Goal: Given two language names, find parallel texts.

- Hervé Saint-Amand’s master’s thesis (Saarbrücken).
  - Train language identification on Wikipedia.
  - Search for pages in English containing the word česky.
- PANACEA tools ([http://myexperiment.elda.org/workflows/7](http://myexperiment.elda.org/workflows/7))
- Students’ project ParaSite: proof of concept, fixes needed.

Quasi-comparable sources (incl. Wikipedia):

- Texts on the same topic but written independently.
- Can hope to find parallel sentences but no longer segments.
- “Lightly supervised training” (Schwenk, 2008) = basis of unsupervised MT.
Goal: Given bag of texts in two languages, find pairs.

- A project at this very seminar at FJFI: (Jahoda et al., 2007)
- A project at MFF: (Klempová et al., 2009)
  - Evaluation suggested that the first step is tricky: finding source URLs.
  - Proper minimum pairing algorithm.
  - Not generic enough: focus on named entities at the beg. and end only.
- ParaSite: probably good, re-evaluation would be useful.
  - Problem: Based on libraries with conflicting licenses (GPL 2.0 vs 3.0).
- Parallel **Paragraphs** from CommonCrawl (Kúdela et al., 2017)
  - Recall 63%, precision 94% when re-aligning shuffled CzEng.
  - 149TB of CommonCrawl $\Rightarrow$ 115k en-cs sentpairs from 2k webdomains.
  - **Targetted re-crawl would be highly desirable (project suggestion).**
- paracrawl.eu large but noisy. Aligns documents, not paragraphs.
Goal: Given a text in two languages, align sentences.
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
In my dream, there was a sycamore growing out of the ruins of the sanctum, 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.

“Is it who dared to do so,” said the boy.

This man looked exactly the same, except that now the roles were reversed.

It was the end of the world. The boy lowered his head to receive a blow from the sword.

“Life was good to me,” the man said.

“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.”

“I don’t want anything for myself.”

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.

And he gave the boy his blessing.

“You’ll learn to love the desert, and you’ll get to know every one of the fifty thousand palms.”

“You’ll watch them as they grow, demonstrating how the world is always changing.”

And you’ll get better and better at understanding omens, because the desert is the best teacher there is.

Sometimes during the second year, you’ll remember about the treasure.

The omens will begin insistently to speak of it, and you’ll try to ignore them.

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.

“Who told you that?” asked the boy, startled.

“Maktab,” said the old crystal merchant.

And he gave the boy his blessing.

The boy went to his room and packed his belongings.

They filled three sacks.

As he was leaving, he saw, in the corner of the room, his old shepherd’s pouch.

“I want to see the greatness of Allah,” the chief said, with respect.

“I want to see how a man turns himself into the wind.”

But he made a mental note of the names of the two men who had expressed their fear.
Sentence Alignment

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

  - 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).

Illustration: MT Talk #7 (https://youtu.be/_4lnyoC3mtQ)
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: \( P(\text{kočka} = \text{cat}) \)
  - IBM2: absolute reordering added (not used in practice now)
  - IBM3: adds fertility: 1 word generates several others
  - IBM4/HMM: to account for relative reordering

- Only many-to-one links created \( \Rightarrow \) used twice, in both directions.
IBM Model 1

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

See the slides by Philipp Koehn for:
- Formulas of both expectation and maximization step.
- The trick in expectation step, swapping sum and product by rearranging the sum.
- Pseudocode.

The Trick Illustrated

Sum of pairs:

Can be rearranged:
EM Loop in IBM1

Illustration from Bojar (2012)

Iteration 0

\[ \begin{array}{cccc}
\text{bílý} & \text{white} & \text{house} & \Sigma \\
\text{dům} & \Sigma & \Sigma & \Sigma \\
\text{bílý} & \text{white} & \text{dog} & \Sigma \\
\text{pes} & \Sigma & \Sigma & \Sigma \\
\text{černý} & \text{black} & \text{dog} & \Sigma \\
\text{pes} & \Sigma & \Sigma & \Sigma \\
\end{array} \]

\[ \begin{array}{cccc}
p(\downarrow | \rightarrow) & \text{black} & \text{dog} & \text{house} & \text{white} \\
\text{bílý} & \Sigma & \Sigma & \Sigma & \Sigma \\
\text{černý} & \Sigma & \Sigma & \Sigma & \Sigma \\
\text{dům} & \Sigma & \Sigma & \Sigma & \Sigma \\
\text{pes} & \Sigma & \Sigma & \Sigma & \Sigma \\
\Sigma & \Sigma & \Sigma & \Sigma & \Sigma \\
\end{array} \]
“Symmetrization” of two GIZA++ runs:
- intersection: high precision, too low recall.
- popular: heuristical (something between intersection and union).
- minimum-weight edge cover (Matusov et al., 2004).
Popular Symmetrization Heuristic

María no daba una bofetada a la bruja verde

Mary  did  not  slap  the  green  witch
Troubles with Word Alignment

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

<table>
<thead>
<tr>
<th>Top Problematic Words</th>
<th>Top Problematic Parts of Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td><strong>Czech</strong></td>
</tr>
<tr>
<td>361 to</td>
<td>319 ,</td>
</tr>
<tr>
<td>259 the</td>
<td>271 se</td>
</tr>
<tr>
<td>159 of</td>
<td>146 v</td>
</tr>
<tr>
<td>143 a</td>
<td>112 na</td>
</tr>
<tr>
<td>124 ,</td>
<td>74 o</td>
</tr>
<tr>
<td>107 be</td>
<td>61 že</td>
</tr>
<tr>
<td>99 it</td>
<td>55 .</td>
</tr>
<tr>
<td>95 that</td>
<td>47 a</td>
</tr>
</tbody>
</table>
## Limits of Automatic W.A.

<table>
<thead>
<tr>
<th>Humans</th>
<th>GIZA++</th>
<th>Baseline</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>cs</td>
</tr>
<tr>
<td>Problems</td>
<td>Problems</td>
<td>14.3</td>
<td>15.5</td>
</tr>
<tr>
<td>Problems</td>
<td>OK</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>OK</td>
<td>Problems</td>
<td>38.6</td>
<td>35.7</td>
</tr>
<tr>
<td>OK</td>
<td>OK</td>
<td>46.9</td>
<td>48.7</td>
</tr>
</tbody>
</table>

Percentage of English (en) and Czech (cs) tokens where the alignment was difficult for humans and/or for GIZA++. (Humans against each other, GIZA++ against merged humans.)

- Where GIZA++ had problems, humans often disagreed, too.
- Improving automatic alignment keeps the problematic part intact.
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

Distribution over possible link types

- 65%
- 31%
A Czech-English Example

Nemyslím, že by se jejich zákazníkům moc líbilo.

I do think would very n't their like much customers .

- Two papers independently published the same work and on the same dataset.
  - Kruijff-Korbayová et al. (2006)
  - Bojar and Prokopová (2006)
- The both defined essentially the same rules.
T-Layer to the Rescue

- Only content-bearing words have a node.
- Auxiliary words hidden, dropped pronouns added.

(já) Nemyslím, že by se jejich zákazníkům moc líbilo.
I do n’t think their customers would like it very much.
Mareček et al. (2008) align t-nodes, not words.
⇒ Auxiliary words do not clutter the task.

Implements human agreement from 91% to 94.7%.

Application to phrase-based MT: (Mareček, 2009)
- Improved alignment error rate on content words.
- Minor improvements in BLEU when combined with GIZA++.

Main use: Extraction of t-lemma dictionaries for e.g. TectoMT.

Main disadvantage:
- Language-dependent.
- Heavy use of tools (tagging, parsing, deep parsing).
Related: Fraser and Marcu (2007)

- A generative story called “LEAF” divides:
  - Source words into classes: head, non-head, deleted.
  - Target words into classes: head, non-head, spurious.
  - Heads connected across languages, non-heads within languages.

- Probabilities in the generative story learnt unsupervised:
  - Starting from GIZA++ outputs.
  - Greedy local updates of alignments to increase the likelihood of the data.

Project suggestions: (1) Revive LEAF, (2) Your own NN version of LEAF.
Phrase extraction based on word alignments is wrong:

- From statistical point of view:
  - No link to the decoding, i.e. the use of the phrases in MT.
  - Wuebker et al. (2010) run “forced” or “constraint” decoding on the training data to obtain phrasal alignments.
  - The overfitting to long phrases is avoided by “leaving-one-out” (Ney et al., 1995).

- From linguistic point of view:
  - Fraser and Marcu (2007) allow for M-to-N non-consecutive translation units.
  - DeNero and Klein (2010) train on manual word alignments and handle “possible” links specifically.
The better (more fluent) translation, the harder to align:

to o * - - - - - - - - - -
get - - * - - - - - - - - -
in - - - - - - - @ 0 0 0 -
shape - - - - - - - 0 0 0 @ -
for - - - * - - - - - - - - -
the - - - - - - - o - - - - -
1990s - - - - * * * - - - - -
. - - - - - - - - - - - *

, aby do . let co formě
vstoupila v nejlepší
90 .
T-layer to no rescue:

Teď se zdá, že tyto trasy začnou fungovat až v lednu.

Now, those routes are n’t expected to begin until Jan.
Summary

- Parallel data are vital for MT. The more and better, the better.
- Several projects for document alignment.
  Project suggestion: Targeted re-crawl based on Kúdela et al. (2017).
- Sentence alignment “solved”.
- Word alignment ill-defined but used to be very important.
  Plus all the funny heuristics…
- Beyond word alignment:
  - Phrase alignment never got wide-spread; too tied to PBMT anyway.
  - T-Alignment costly (T-layer needed).
  - Project suggestion: NN LEAF.
References


Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. 2008. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), Marrakech, Morocco, May. ELRA.


Ondřej Bojar. 2013. Čeština a strojový překlad (Czech Language and Machine Translation) – volume 11 of Studies in