Alignment

Ondřej Bojar

March 19, 2020

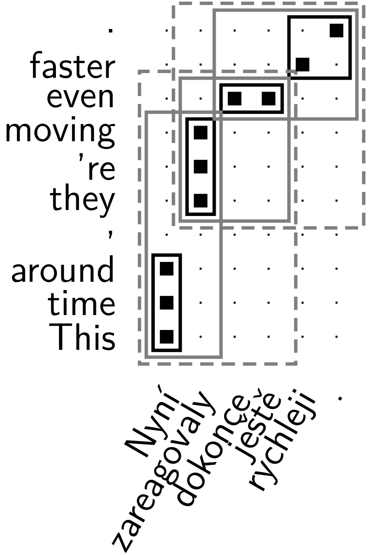
EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

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

Outline

- 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



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

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

	Sentences	Tokens
Acquis Communautaire	64.1%	69.0%
Readers' Digest	8.6%	8.6%
Project Syndicate	6.5%	8.9%
KDE Messages	**6.2%**	**1.9%**
GNOME Messages	**5.7%**	**1.9%**
Kačenka	4.2%	4.9%
Navajo User Translations	**2.3%**	**2.1%**
E-Books	1.2%	1.6%
European Constitution	0.8%	0.7%
Samples from European Journal	0.4%	0.5%
Total	1.4 mil.	21 mil.

Community-supplied data in bold.

Community-Supplied Data (1/2)

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:

Translation Quality	Proportion in the Sample
precise, flawless	**69.0%**
not translated	6.8%
incomplete	6.6%
imprecise	5.8%
precise, almost flawless	**4.5%**
machine-generated	4.4%
vandalism	2.7%
other	0.2%

Community-Supplied Data (2/2)

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”

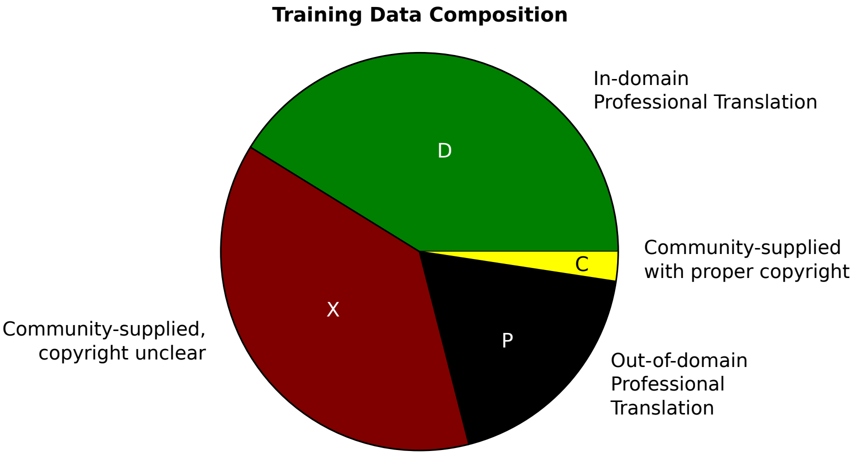
Licensing Issues

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

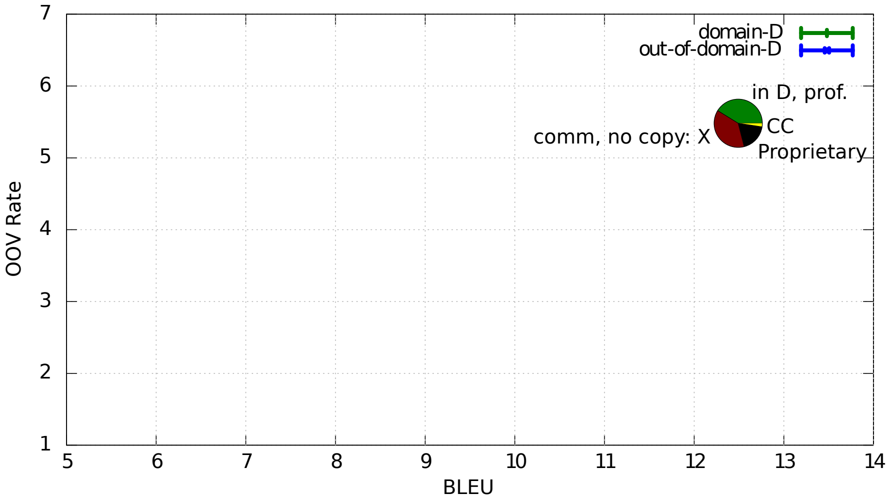
	Tokens		Available	
Source of Texts and Translation	cs	en	cs	en
Community Transl. of Proprietary Texts	19.5M	25.3M	37.8%	41.1%
Professional	21.3M	23.9M	41.2%	38.9%
Proprietary	9.6M	10.9M	18.6%	17.7%
Community	1.2M	1.4M	2.4%	2.3%
Total	51.6M	61.5M	100.0%	100.0%

CzEng 0.7 $\approx$ Professional + Community sources; in bold

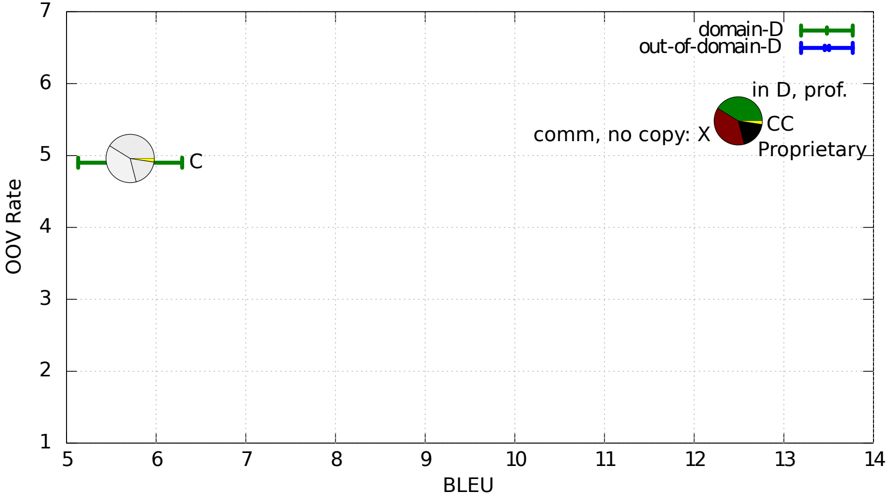
En→Cs Data in 2008



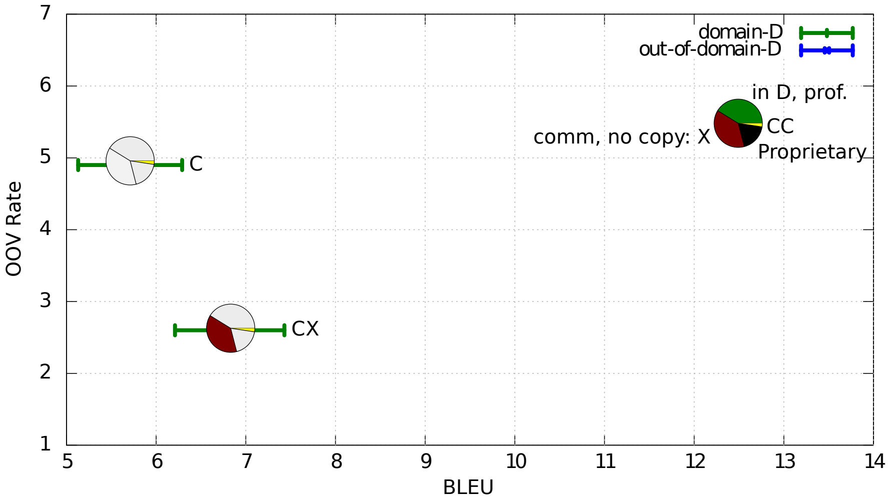
OOV and PBMT Quality In/Out of Domain



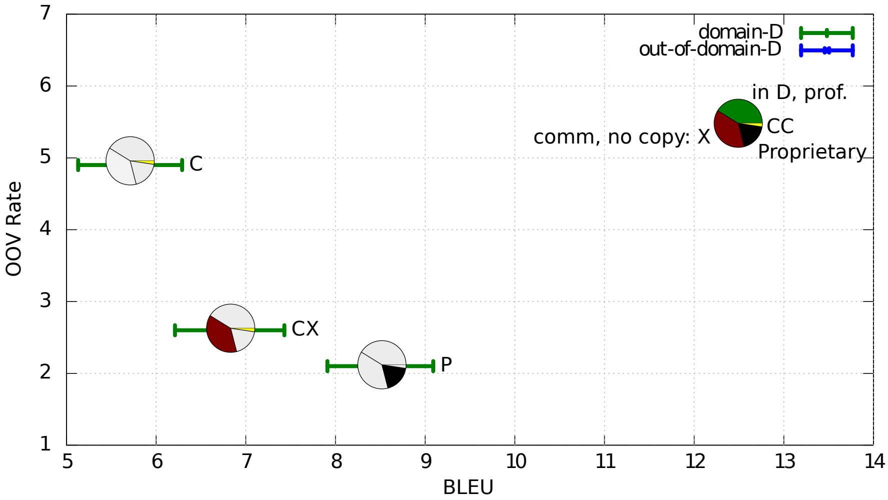
Community Data Out-of-Domain



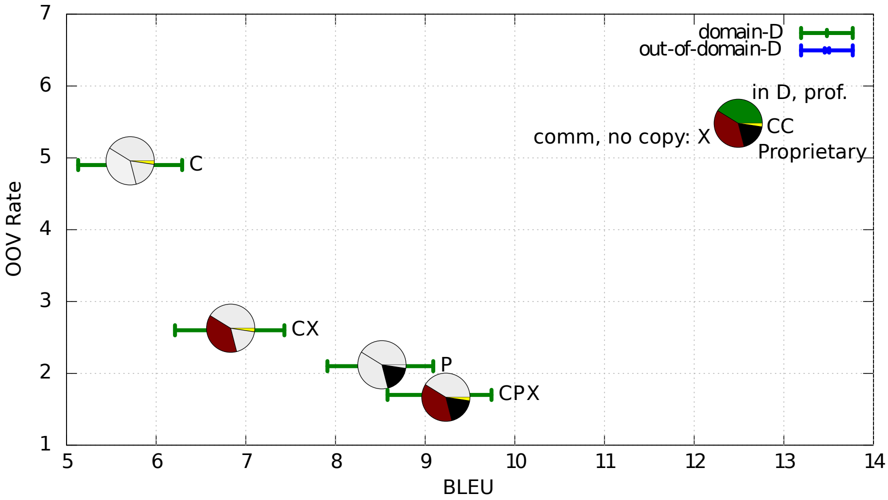
Community Data Out-of-Domain



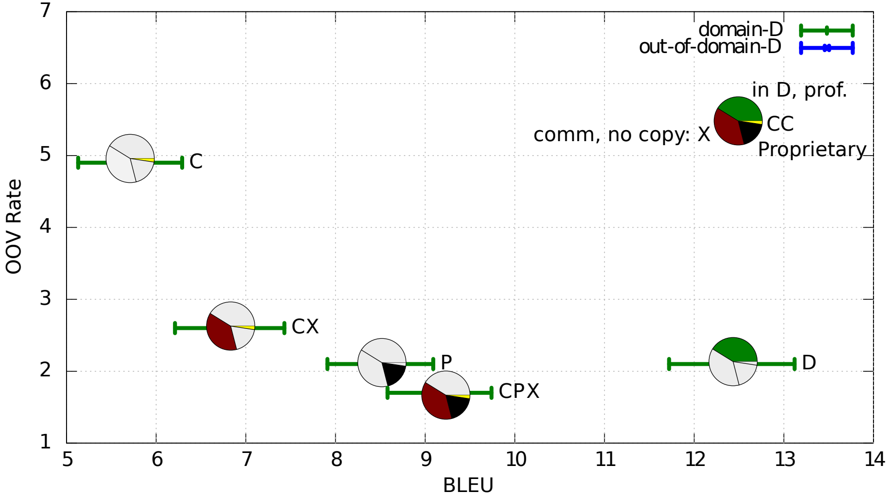
Professional Out-of-Domain



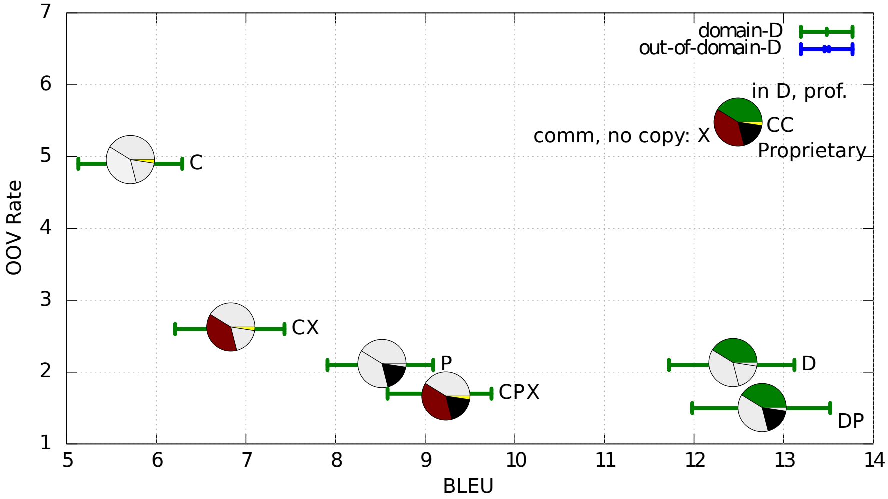
Everything Out-of-Domain



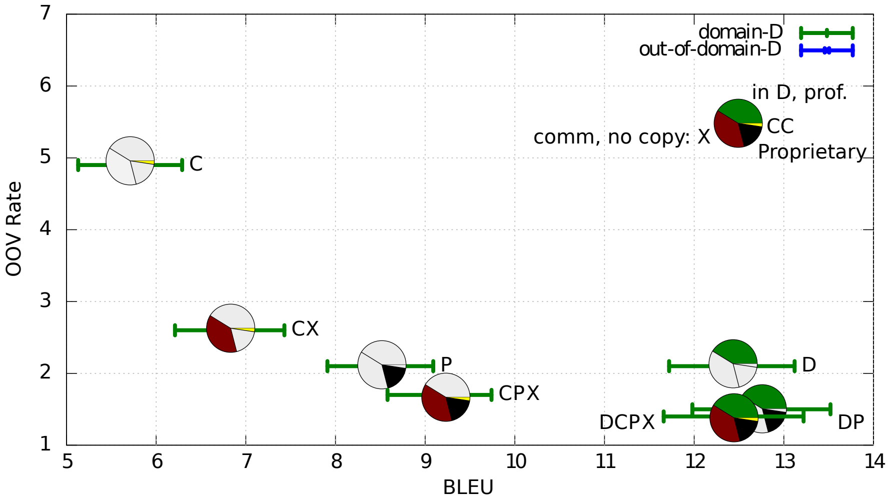
Similar Volume of in-Domain: Much Better



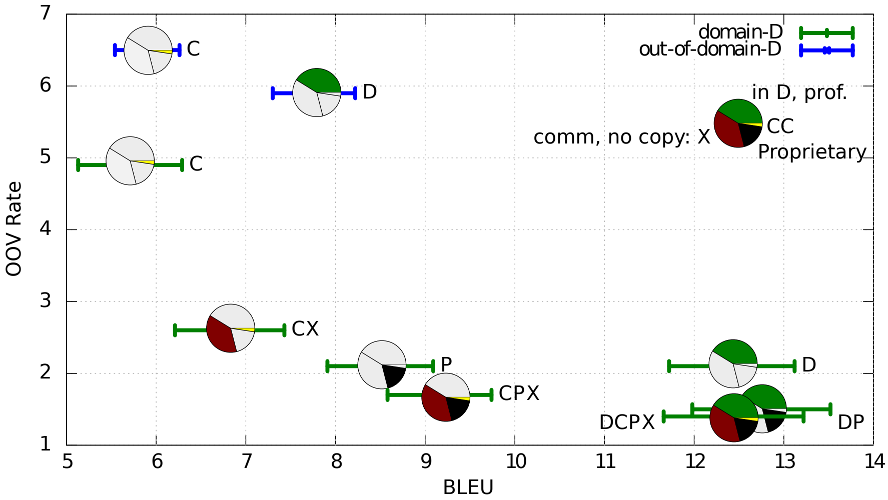
Additional Data Improve Coverage



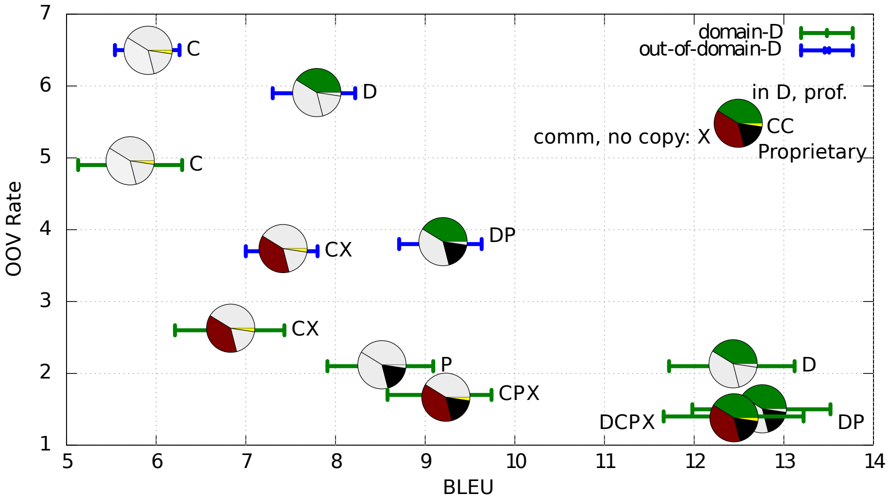
But Out-of-Domain Can Decrease Quality



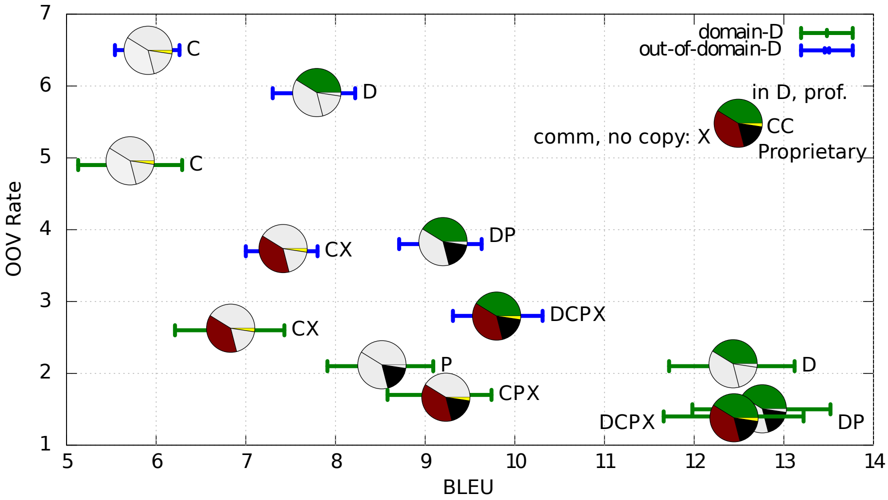
Applying Out of Domain? Much Worse.



More Data → Better Coverage



...But Not Much Better Quality



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)

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

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*.
- Bitextor: Esplà-Gomis and Forcada (2010)
- PANACEA tools (`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.
- BUCC workshops 2008–2020: `https://comparable.limsi.fr/bucc2020/`
- "Lightly supervised training" (Schwenk, 2008) = basis of unsupervised MT.

Document Alignment Attempted Many Times

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.
- Václav Novák (ÚFAL, ∼2009): aligning subtitles.
    - 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 ⇝ 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.

Sentence Alignment

Goal: Given a text in two languages, align sentences.

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).

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: $P(\textit{kočka} = \textit{cat})$
  - 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.

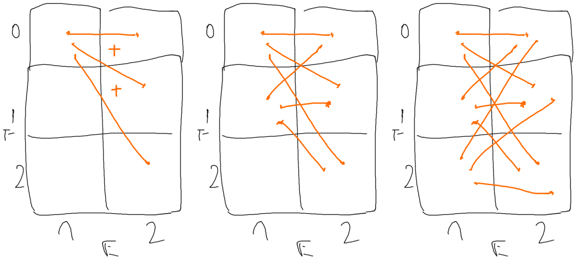
See the slides by Philipp Koehn for:

- Formulas of both expectation and maximization step.
- The trick in expectation step, swapping sum and product by rearranging the sum.
- Pseudocode.

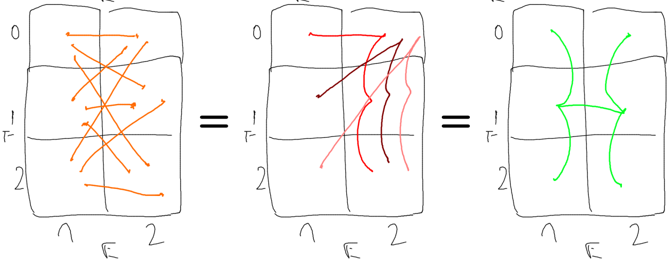
Illustration: MT Talk #8 (`https://youtu.be/mqyMDLu5JPw`)

The Trick Illustrated

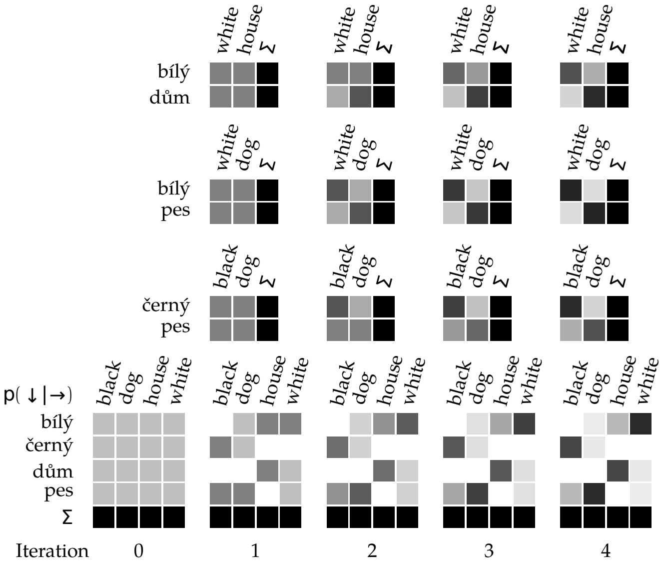
Sum of pairs:



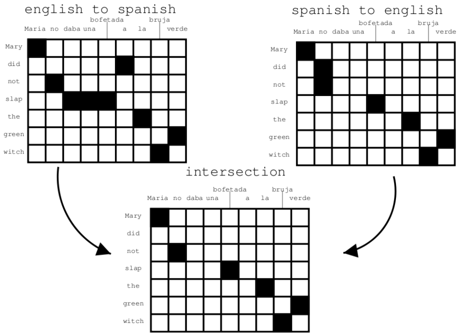
Can be rearranged:



EM Loop in IBM1 Illustration from Bojar (2012)



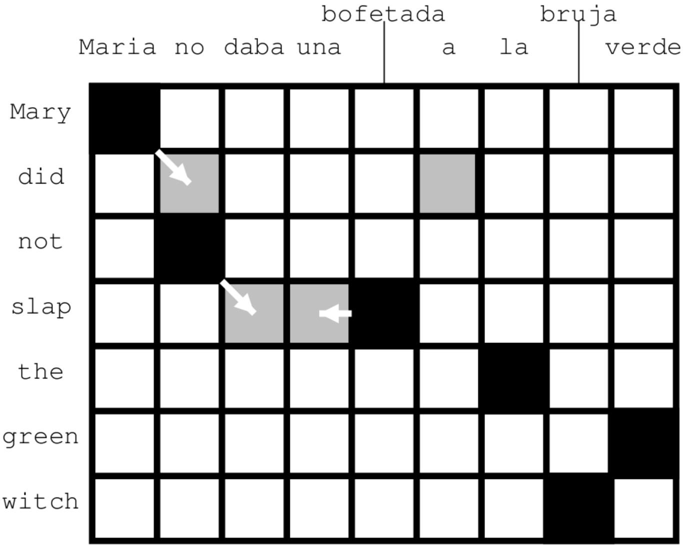
Symmetrization



"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



Troubles with Word Alignment

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

Top Problematic Words

English		Czech	
361	to	319	,
259	the	271	se
159	of	146	v
143	a	112	na
124	,	74	o
107	be	61	že
99	it	55	.
95	that	47	a

Top Problematic Parts of Speech

English		Czech	
679	IN	1348	N
519	DT	1283	V
510	NN	661	R
386	PRP	505	P
361	TO	448	Z
327	VB	398	A
310	JJ	280	D
245	RB	192	J

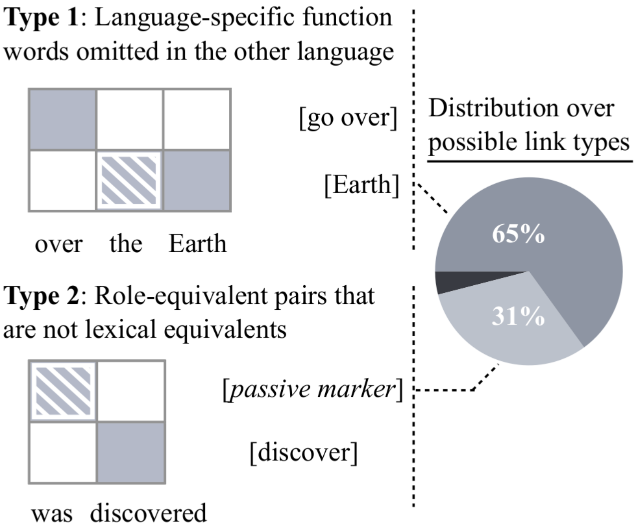
Limits of Automatic W.A.

		Baseline		Improved	
Humans	GIZA++	en	cs	en	cs
Problems	Problems	14.3	15.5	14.3	15.5
Problems	OK	0.1	0.1	0.2	0.1
OK	Problems	38.6	35.7	25.2	25.0
OK	OK	46.9	48.7	60.4	59.4

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



Chinese-English from DeNero and Klein (2010).

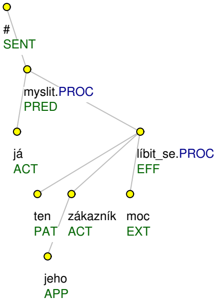
A Czech-English Example

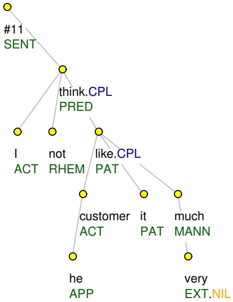
	I	do	n't	think	their	customers	would	like	it	very	much	.
Nemyslím	o	o	o	*	-	-	-	-	-	-	-	-
,	-	-	-	-	-	-	-	o	-	-	-	-
že	-	-	-	-	-	-	-	o	-	-	-	-
by	-	-	-	-	-	-	-	o	-	-	-	-
se	-	-	-	-	-	-	-	o	-	-	-	-
to	-	-	-	-	-	-	-	-	*	-	-	-
jejich	-	-	-	-	*	-	-	-	-	-	-	-
zákazníkům	-	-	-	-	-	*	-	-	-	-	-	-
moc	-	-	-	-	-	-	-	-	-	*	*	-
líbilo	-	-	-	-	-	-	-	*	-	-	-	-
.	-	-	-	-	-	-	-	-	-	-	-	*

- 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** to jejich zákazníkům moc líbilo **.**

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

Tectogrammatical Alignment

- Mareček et al. (2008) align t-nodes, not words.
  $\Rightarrow$ Auxiliary words do not clutter the task.
- Improves 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.

source	absolutely	[comma]	they	do	not	want	to	spend	that	money
word type (1)	DEL.	DEL.	HEAD	non-head	HEAD	HEAD	non-head	HEAD	HEAD	HEAD
linked from (2)			THEY	do	NOT	WANT	to	SPEND	THAT	MONEY
head(3)			ILS		PAS	DESIRENT		DEPENSER	CET	ARGENT
cept size(4)			1		2	1		1	1	1
num spurious(5)	1									
spurious(6)	aujourd'hui									
non-head(7)			ILS	PAS	ne	DESIRENT	DEPENSER	CET	ARGENT	
placement(8)	aujourd'hui		ILS	ne	DESIRENT	PAS	DEPENSER	CET	ARGENT	
spur. placement(9)			ILS	ne	DESIRENT	PAS	DEPENSER	CET	ARGENT	aujourd'hui

- 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.

Using Alignment in PBMT

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.

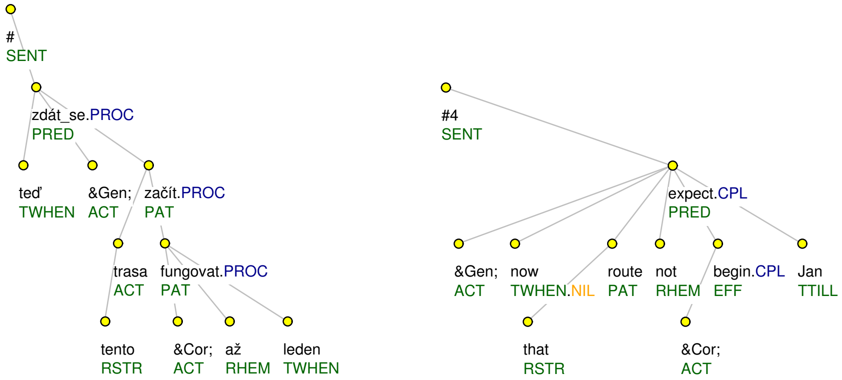
Better Translation ⇝ Uglier Ali. (1)

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            .
```

Better Translation ⇝ Uglier Ali. (2)

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

- Paralel 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 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.

Ondřej Bojar and Zdeněk Žabokrtský. 2006. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng 0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92:63–83.

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, Adam Liška, and Zdeněk Žabokrtský. 2010. Evaluating Utility of Data Sources in a Large Parallel Czech-English Corpus CzEng 0.9. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 447–452, Valletta, Malta, May. ELRA, European Language Resources Association.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3921–3928, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.