Related to: Machine Translation

Vít Baisa

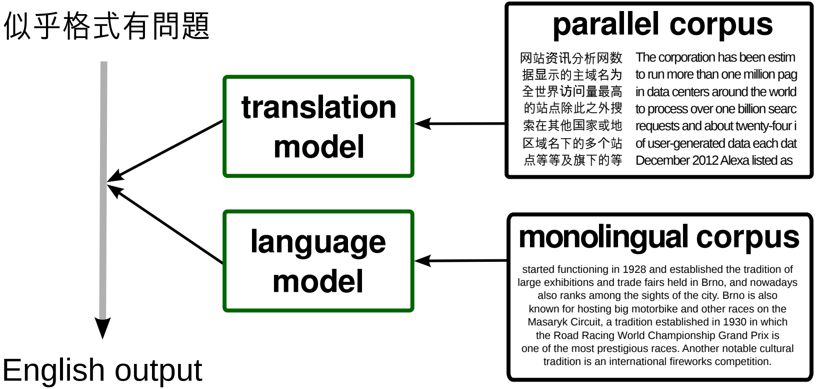
Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

xbaisa@fi.muni.cz

Topics

- Extension of translation memories
- Domain-specific machine translation
- Sub-word level NLP
- Multilingual terminology extraction

Statistical machine translation



似乎格式有問題

translation model

language model

English output

parallel corpus

网站资讯分析网数
据显示的主域名为
全世界访问量最高
的站点除此之外搜
索在其他国家或地
区域名下的多个站
点等等及旗下的等

The corporation has been estim
to run more than one million pag
in data centers around the world
to process over one billion searc
requests and about twenty-four i
of user-generated data each dat
December 2012 Alexa listed as

monolingual corpus

started functioning in 1928 and established the tradition of large exhibitions and trade fairs held in Brno, and nowadays also ranks among the sights of the city. Brno is also known for hosting big motorbike and other races on the Masaryk Circuit, a tradition established in 1930 in which the Road Racing World Championship Grand Prix is one of the most prestigious races. Another notable cultural tradition is an international fireworks competition.

Translation memories

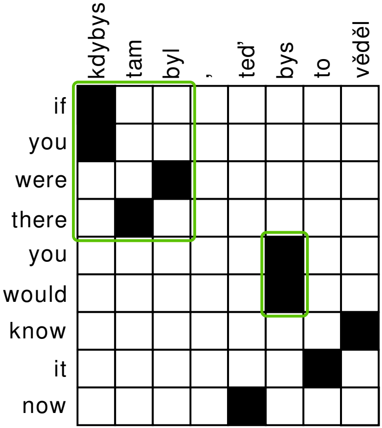
- used in computer-aided translation systems,
- manually built,
- relatively small and focused,
- usually in-house and not for (even academical) use.
- Goal: expand a TM to increase its coverage.
- En↔Cs language pair.

Word alignment matrix – from words to phrases I

	kdybys	tam	byl	,	teď	bys	to	věděl
if	■							
you	■							
were			■					
there		■						
you						■		
would						■		
know								■
it							■	
now					■			

Straightforward utilization for Computer-assisted translation →

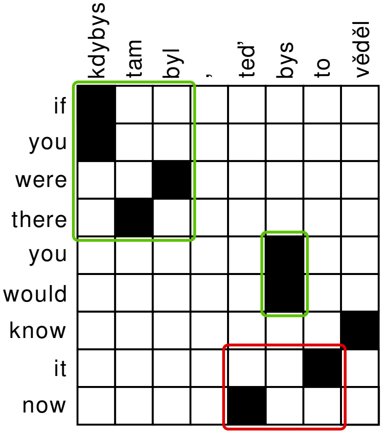
Word alignment matrix – from words to phrases II



	kdybys	tam	byl	,	teď	bys	to	věděl
if	■							
you	■							
were			■					
there		■						
you						■		
would						■		
know								■
it							■	
now					■			

→ Generating new segments in translation memories

Word alignment matrix – from words to phrases III



	kdybys	tam	byl	,	teď	bys	to	věděl
if	■							
you	■							
were			■					
there		■						
you						■		
would						■		
know								■
it							■	
now					■			

→ Generating new segments in translation memories

Evaluation: subsegments generation & combination

We used a sample of TM and a testing document provided by a Czech translation services provider; as evaluation metrics we used the one used by MemoQ (CAT system).

	^{s}TM		$_{sub}$TM		^{s}TM+$_{sub}$TM	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	576	6.4	1247	**15.67**	1286	**17.01**

	^{s}TM		$_{subjoin}$TM		^{s}TM+$_{subjoin}$TM	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	576	6.4	1917	**40.47**	1941	**40.89**

Machine translation of subsegments, example

A sentence from MT:
Návod na použití desinfekčního přípravku najdete na konci této brožury

A manual translation:
You can find instructions for use of disinfectant at the end of this brochure

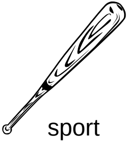
A sentence for translation:
Návod na použití kartáče na vlasy najdete na konci této brožury

Not in TM: *kartáče na vlasy*
Google Translate returns: *hairbrush* (after lemmatization).

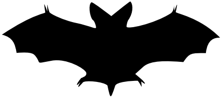
→ Substitute the translation in the existing segment from TM.

Domain-specific machine translation

- straightforward way of increasing quality of MT
- domain-specific corpora can be downloaded on demand
- separate models for each domain: sports, cooking, gardening
- one sense per domain: **bat**



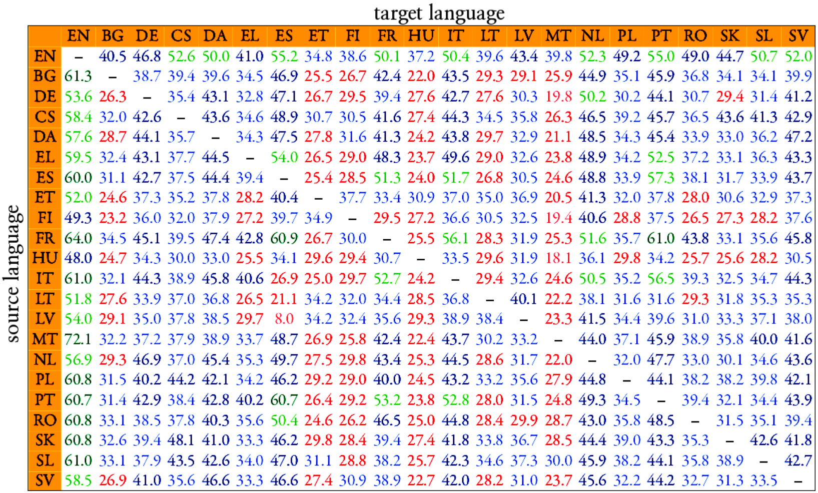
sport



biology

- translations of
  - product details, product descriptions in e-shops,
  - manuals, warranty certificates,
  - user interface localizations, ...

MT quality, European languages



target language

source language

	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Sub-word level machine translation

- SMT principle applied on character level
- translation on subword level (English → Czech)
  -ed → -án, -al, -aný; ex- → vy-
  worked → dělal
  exhausted → vyčerpaný
- translation across levels
  with → -em; user → -ák
  with knife → nožem
  linux user → linuxák

Sub-word level: other possible advantages

- experiments with PoS tagging
- -á, -lá, -alá, -malá
- lyžiny X ližiny (typos)
- edit distance
- language modelling

Multilingual terminology extraction

- input: examined parallel corpus for A ⇔ B; reference corpora for A, B; terminology grammar for A, B
- output: statistically significant keywords/terms, sorted by parallel corpus co-occurence statistics

prevalence	prévalence
soap	savon
survival	survie
education	éducation
primary prevention	prévention primaire
condom	préservatif
chronological age	âge chronologique
basic information	informations de base
acid	acide
universal access	accès universel
international guidance	directives internationales
stigma	stigmatisation
fish	poisson
pregnancy	grossesse
alcohol	alcool
public health	santé publique
disability	handicap
secondary school age	pourcentage du nombre total
training	formation
unemployment	chômage
access	accès
physical appearance	apparence physique
percentage of injecting drug	injection stérile