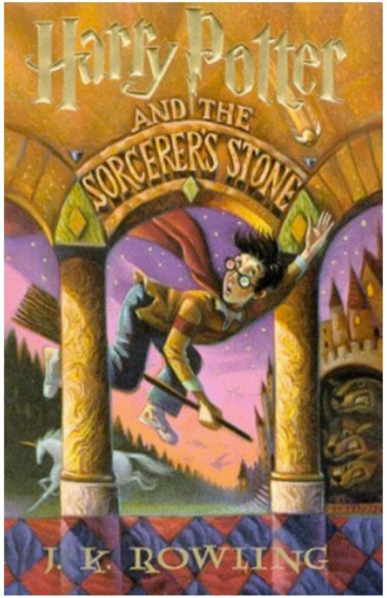
CS11-711: Algorithms for NLP

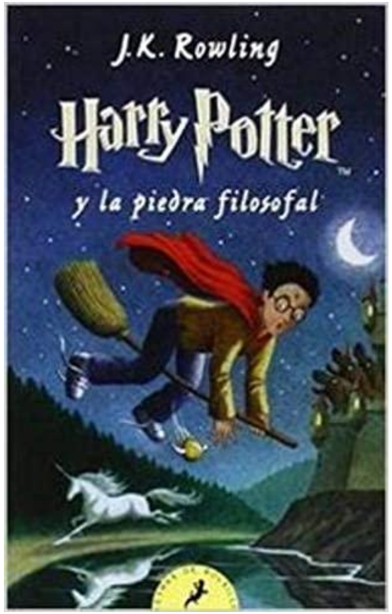
Machine Translation

Yulia Tsvetkov

Carnegie Mellon University
Language Technologies Institute

Translation





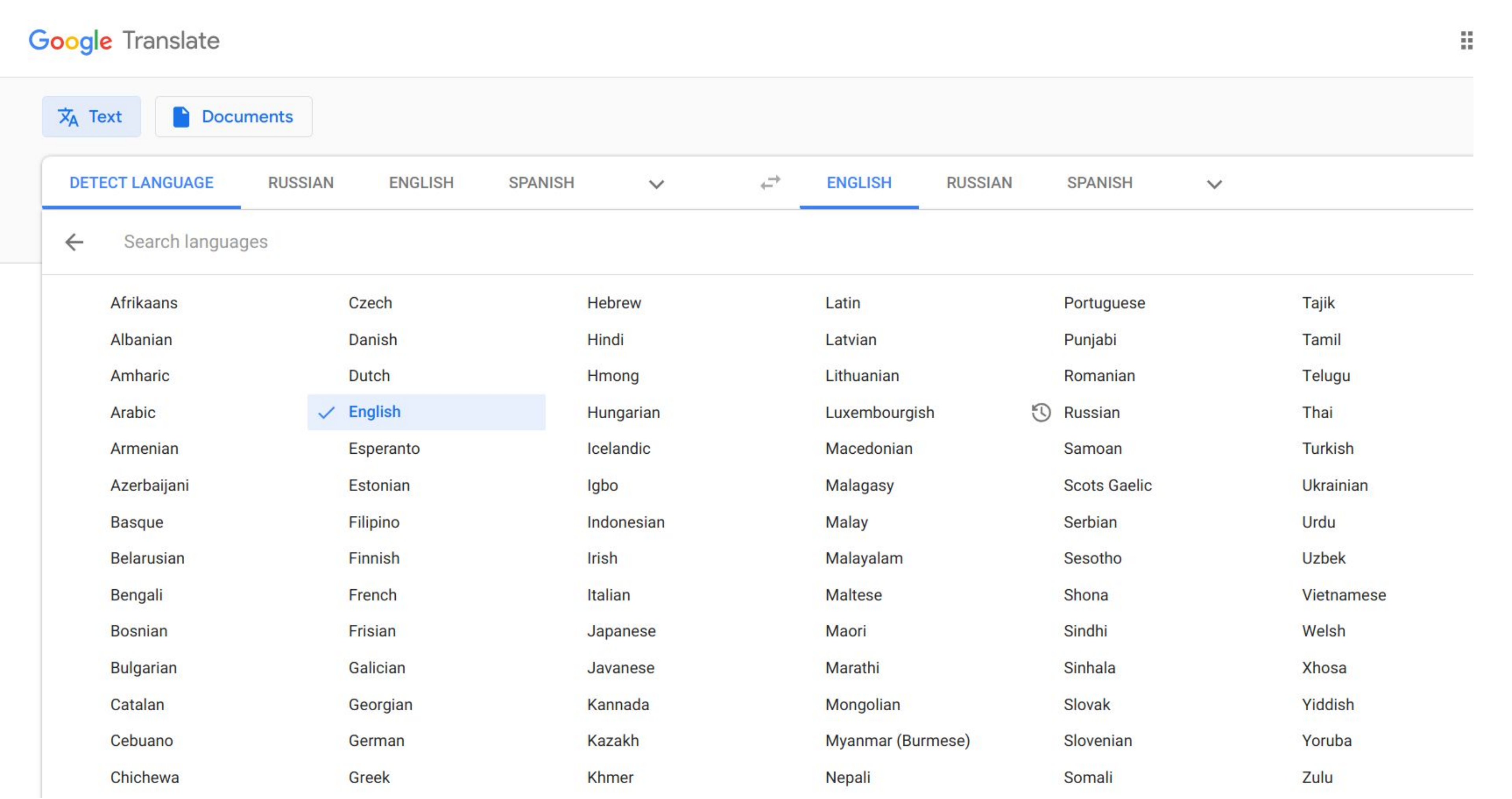
Mr. and Mrs. Dursley, who lived at number 4 on Privet Drive, were proud to say they were very normal, fortunately.

El señor y la señora Dursley, que vivían en el número 4 de Privet Drive, estaban orgullosos de decir que eran muy normales, afortunadamente.

Plan

- The practice of translation
- Machine translation (MT)
- MT data sources
- Modeling: the two views of MT
- MT evaluation

Translation is important and ubiquitous


Google Translate
Text
Documents
DETECT LANGUAGE
RUSSIAN
ENGLISH
SPANISH
ENGLISH
RUSSIAN
SPANISH
Search languages
Afrikaans
Albanian
Amharic
Arabic
Armenian
Azerbaijani
Basque
Belarusian
Bengali
Bosnian
Bulgarian
Catalan
Cebuano
Chichewa
Czech
Danish
Dutch
English
Esperanto
Estonian
Filipino
Finnish
French
Frisian
Galician
Georgian
German
Greek
Hebrew
Hindi
Hmong
Hungarian
Icelandic
Igbo
Indonesian
Irish
Italian
Japanese
Javanese
Kannada
Kazakh
Khmer
Latin
Latvian
Lithuanian
Luxembourgish
Macedonian
Malagasy
Malay
Malayalam
Maltese
Maori
Marathi
Mongolian
Myanmar (Burmese)
Nepali
Portuguese
Punjabi
Romanian
Russian
Samoan
Scots Gaelic
Serbian
Sesotho
Shona
Sindhi
Sinhala
Slovak
Slovenian
Somali
Tajik
Tamil
Telugu
Thai
Turkish
Ukrainian
Urdu
Uzbek
Vietnamese
Welsh
Xhosa
Yiddish
Yoruba
Zulu

Article | Open Access | Published: 01 September 2020

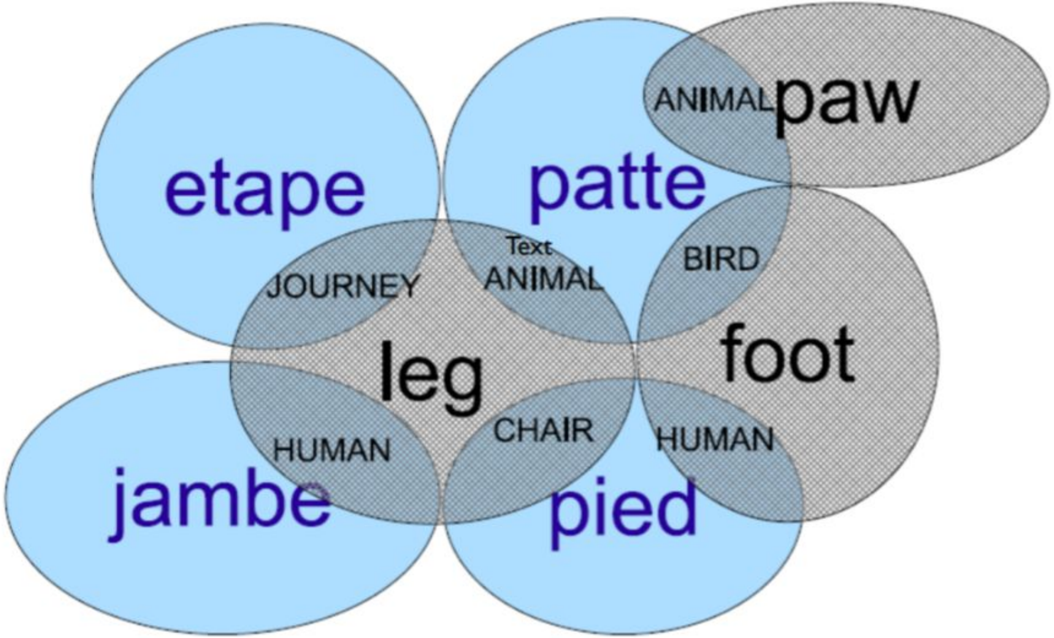
Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar & Zdeněk Žabokrtský

Nature Communications **11**, Article number: 4381 (2020) | Cite this article

Why is it difficult to translate?

- Lexical ambiguities and divergences across languages



[Examples from Jurafsky & Martin Speech and Language Processing 2nd ed.]

Why is it difficult to translate?

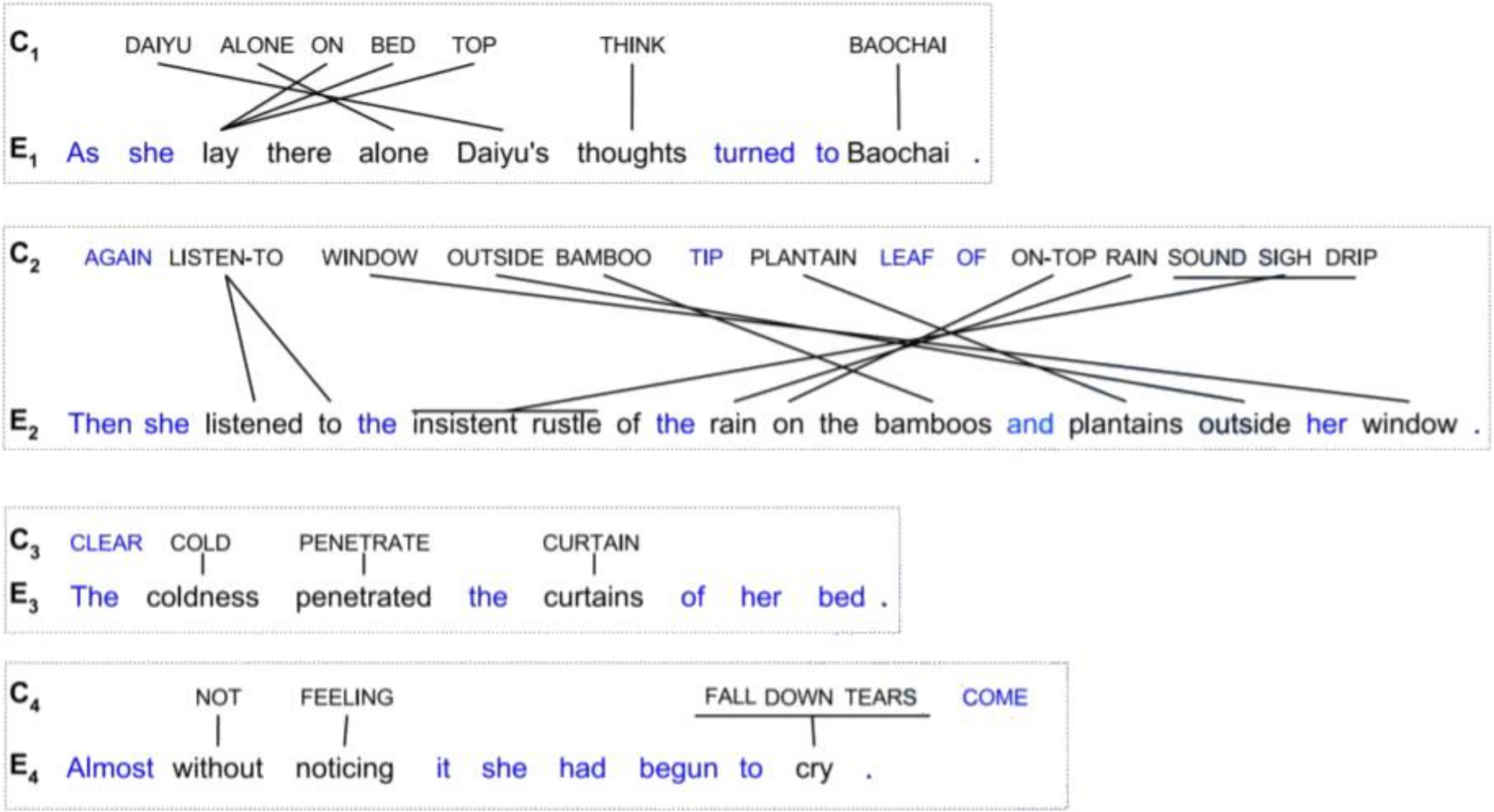
- Cross-lingual lexical and structural divergences

鍸玉自在枕上感念寶釵。。。又聽見窗外竹梢焦葉之上,
雨聲漸沂, 清寒透幕, 不党又滴下淚來 。

dai yu zi zai zhen shang gan nian bao chai...you ting jian chuang wal zhu shao xiang ye
zhe shang, yu sheng xili, qing han tou mu, bu jue you di xia lei lat

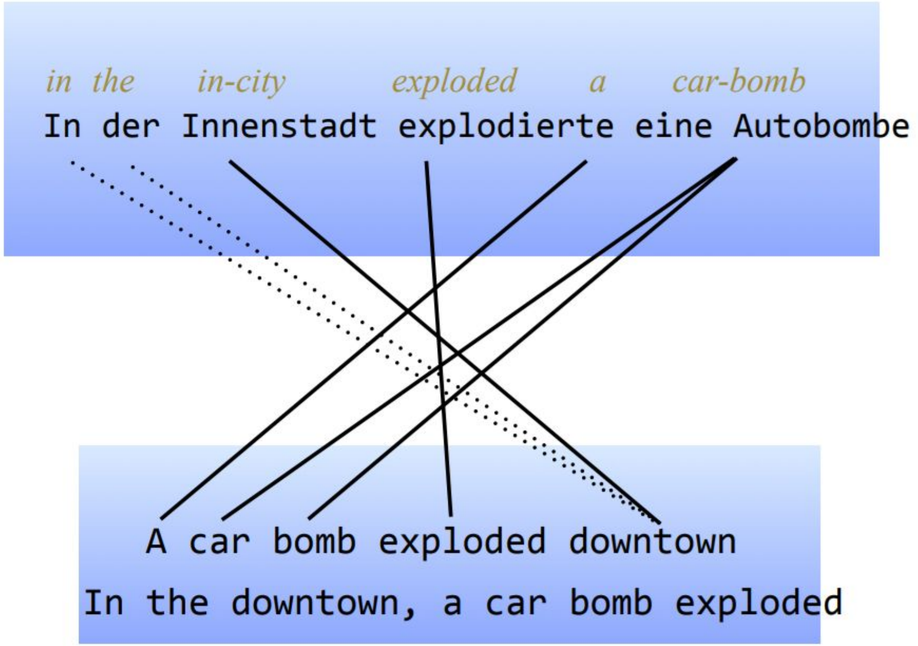
From “Dream of the Red Chamber” Cao Xue Qin (1792)

Why is it difficult to translate?



Why is it difficult to translate?

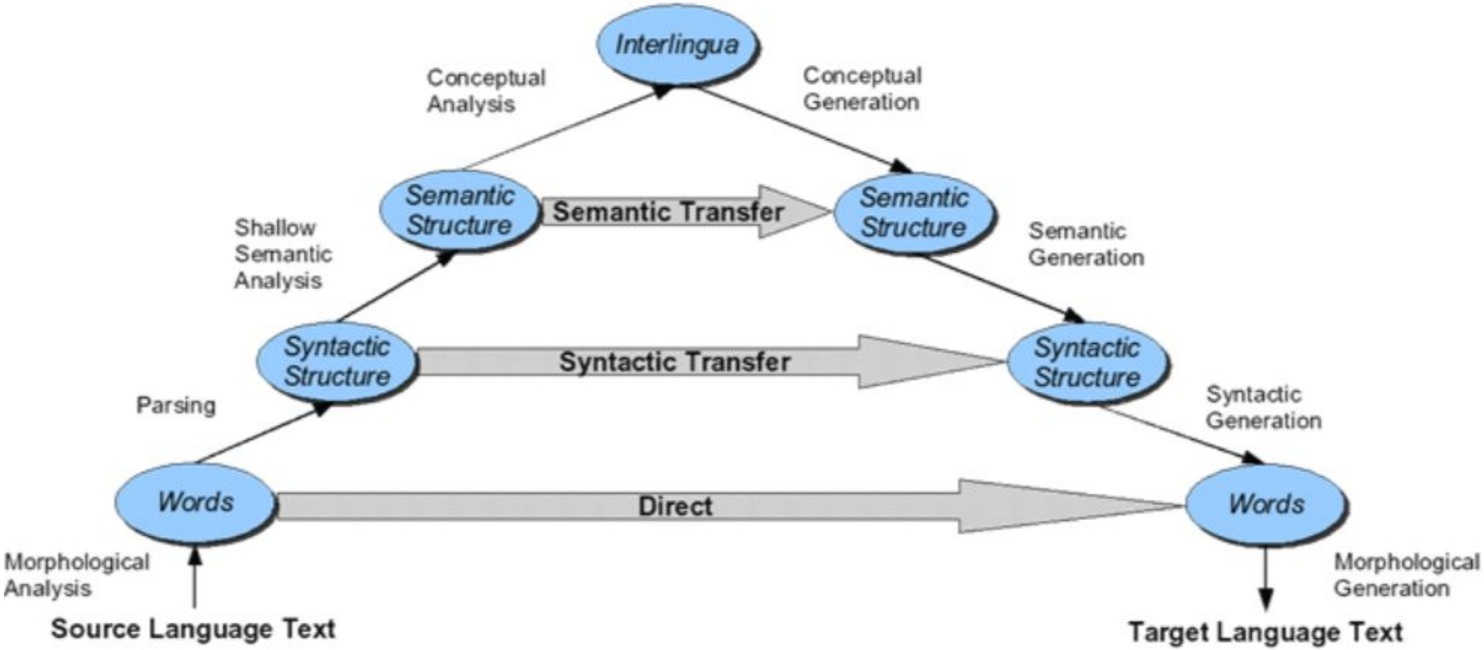
- Ambiguities
  - words
  - morphology
  - semantics
  - pragmatics
- Gaps in data
  - availability of corpora
  - commonsense knowledge
- +Understanding of context,
  connotation, social norms, etc.



3 Classical methods for MT

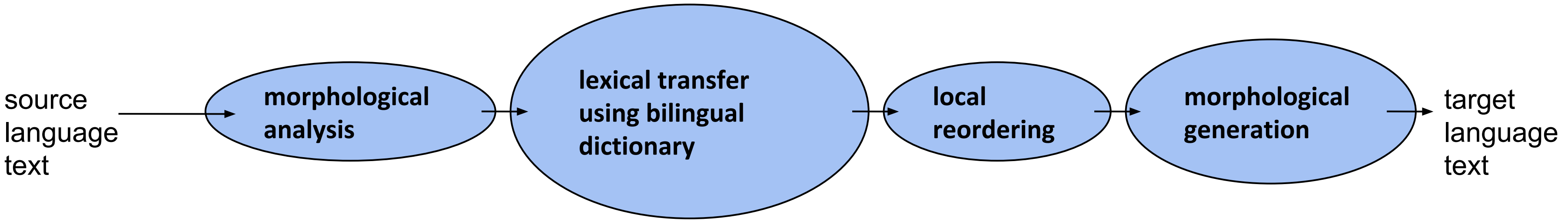
- Direct
- Transfer
- Interlingua

The Vauquois triangle (1968)



Direct translation

- Word-by-word dictionary translation
- Rely on linguistic knowledge for simple reordering or morphological processing



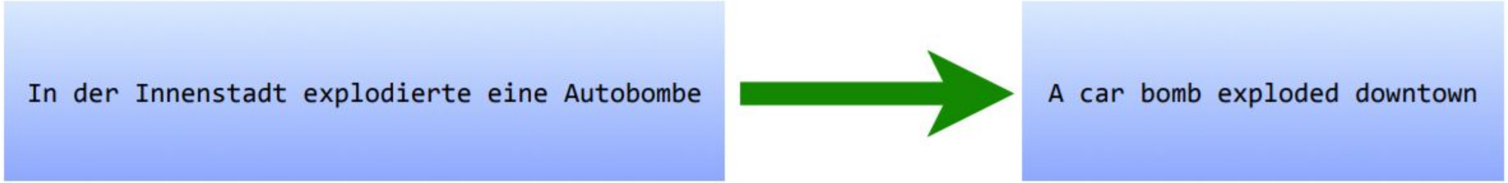
Direct MT dictionary entry

```
function DIRECT_TRANSLATE_MUCH/MANY(word) returns Russian translation

if preceding word is how return skol'ko
else if preceding word is as return stol'ko zhe
else if word is much
    if preceding word is very return nil
    else if following word is a noun return mnogo
else /* word is many */
    if preceding word is a preposition and following word is a noun return mnogii
    else return mnogo
```

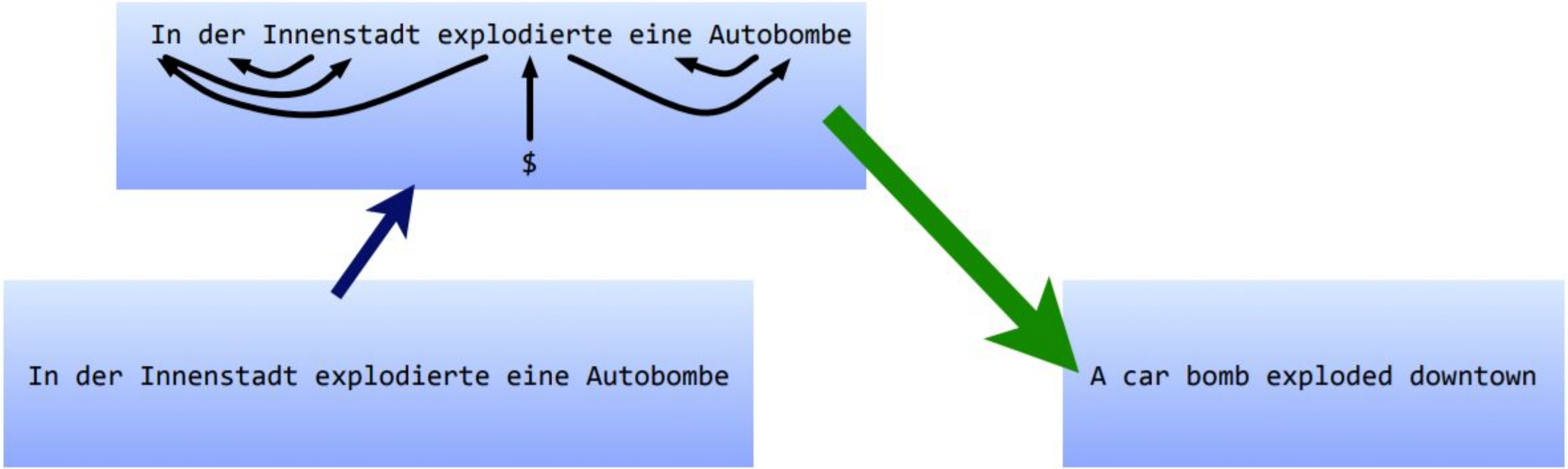
Transfer approaches

- Levels of transfer



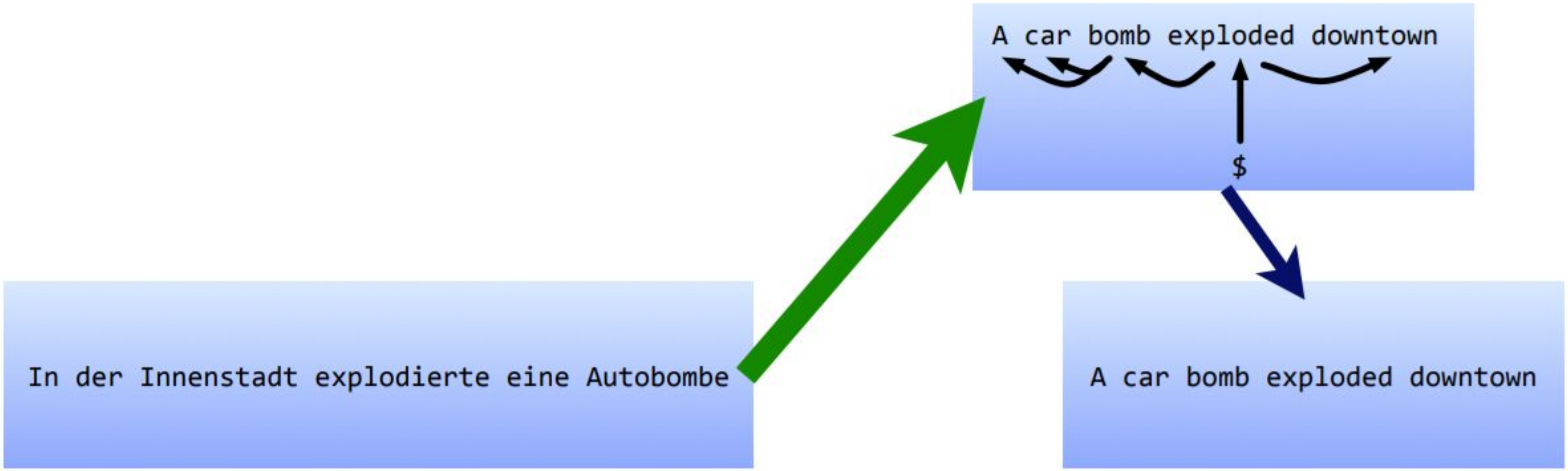
Transfer approaches

- Syntactic transfer



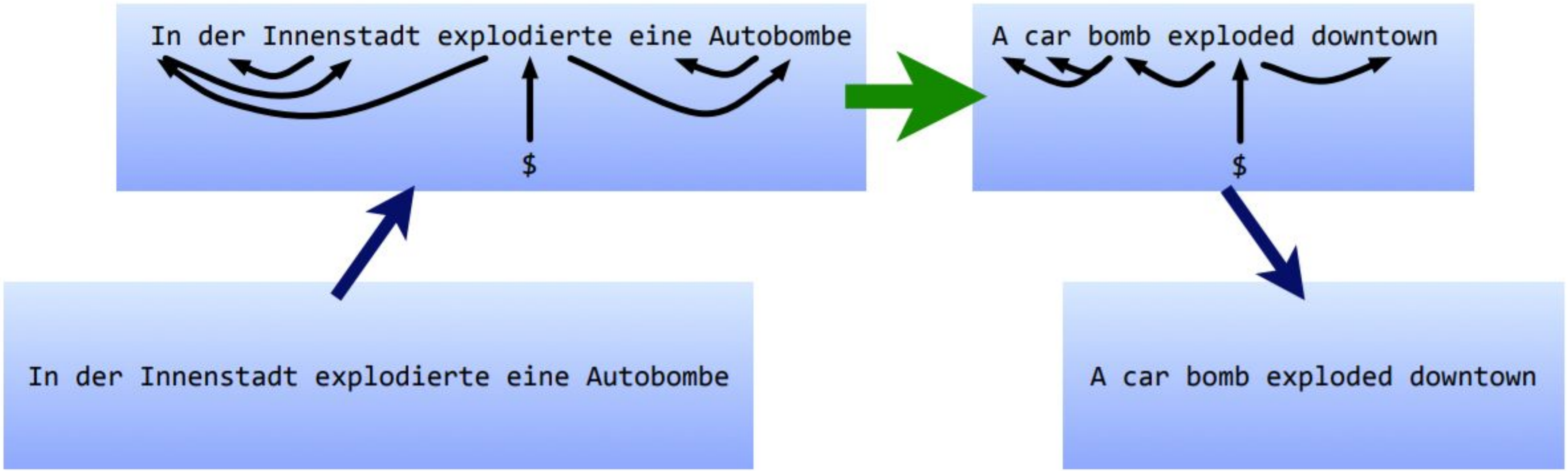
Transfer approaches

- Syntactic transfer



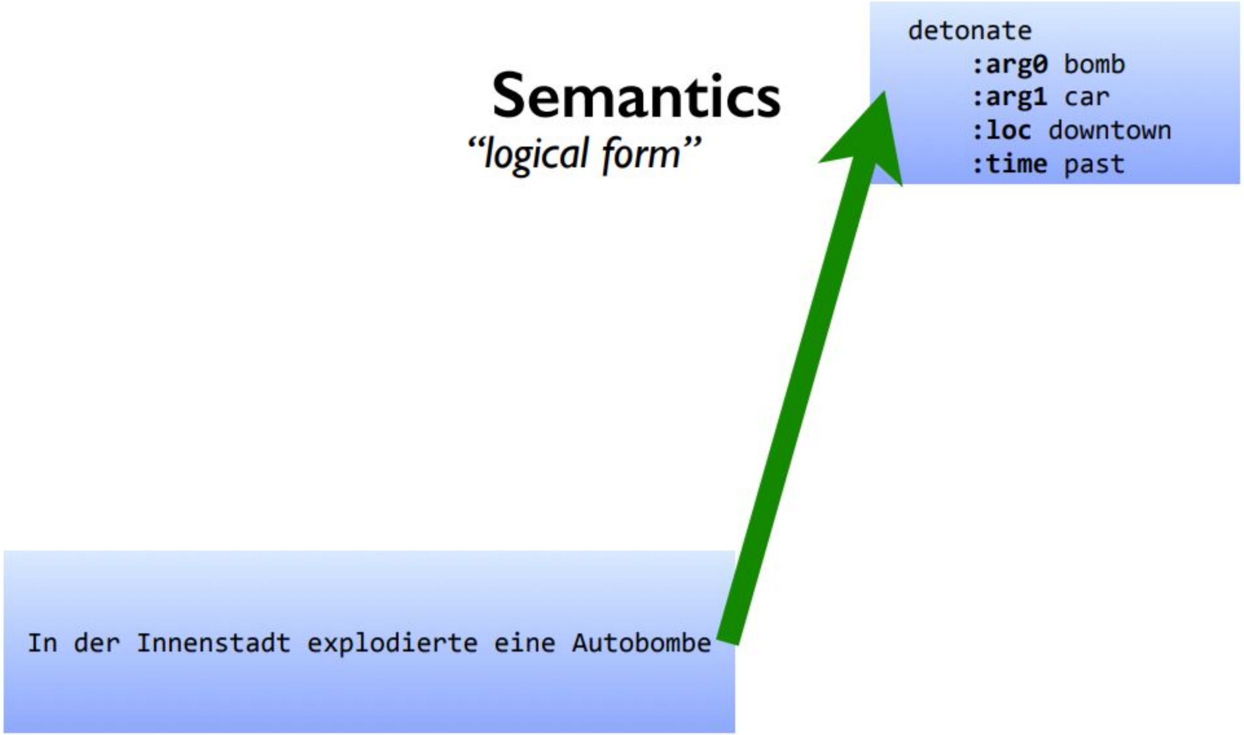
Transfer approaches

- Syntactic transfer



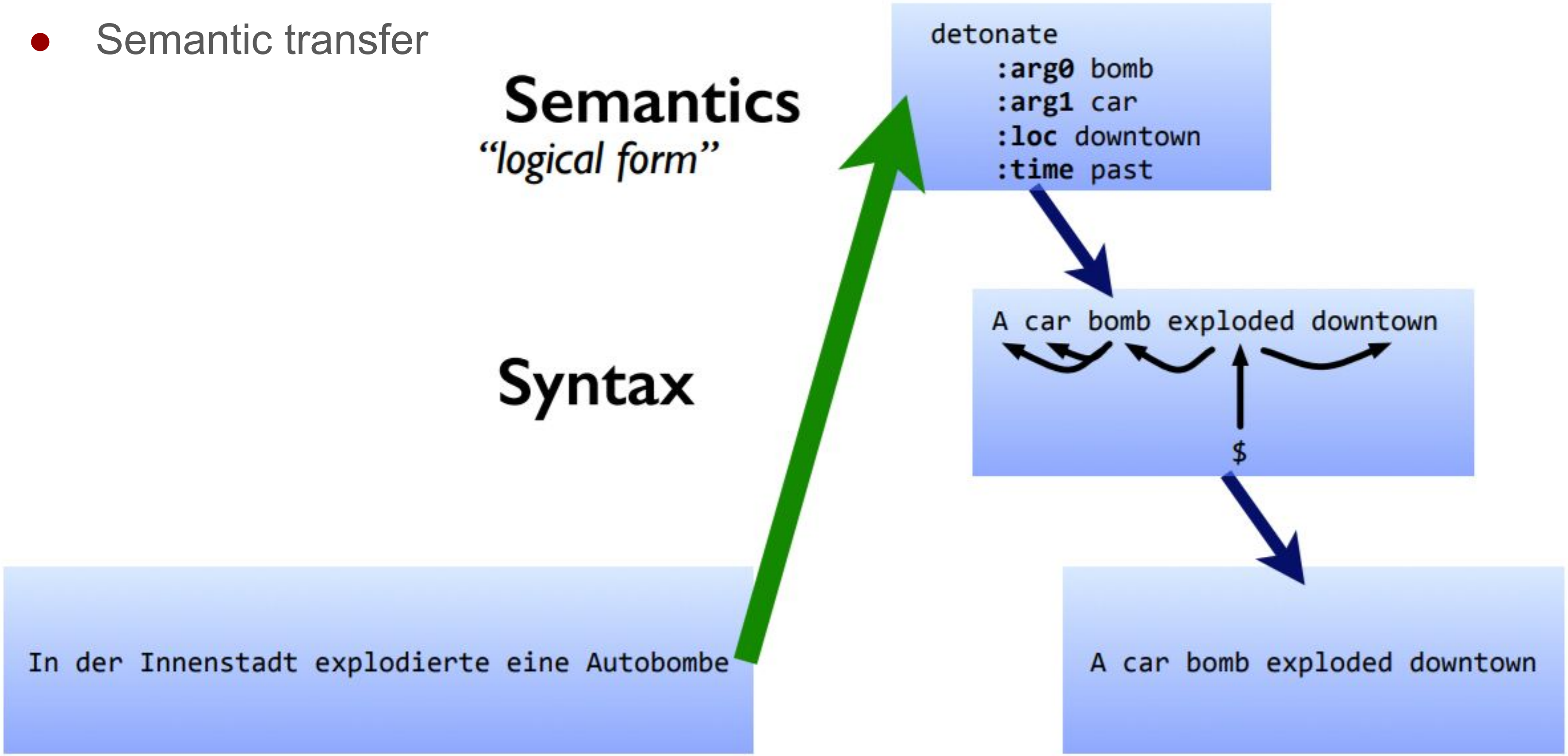
Transfer approaches

- Semantic transfer

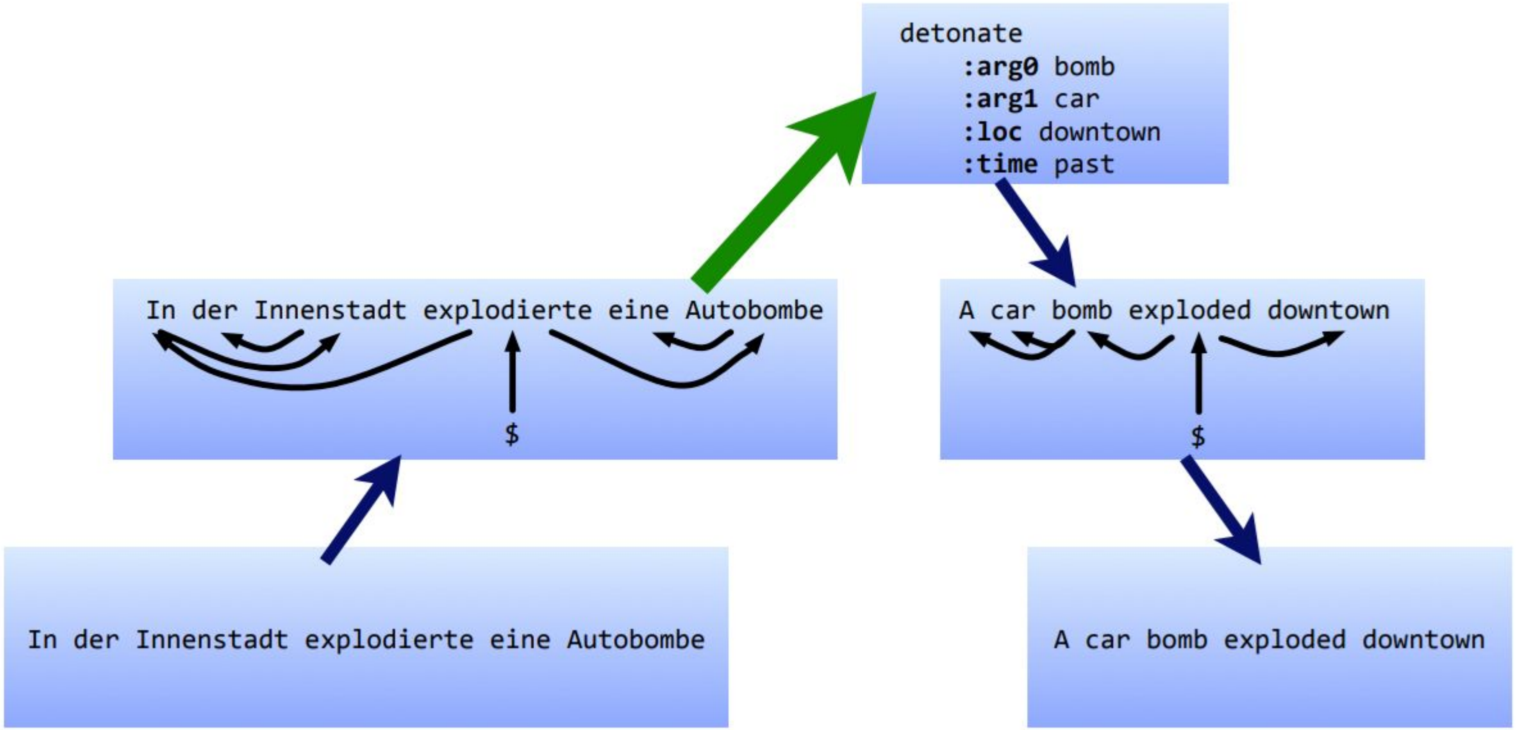


Transfer approaches

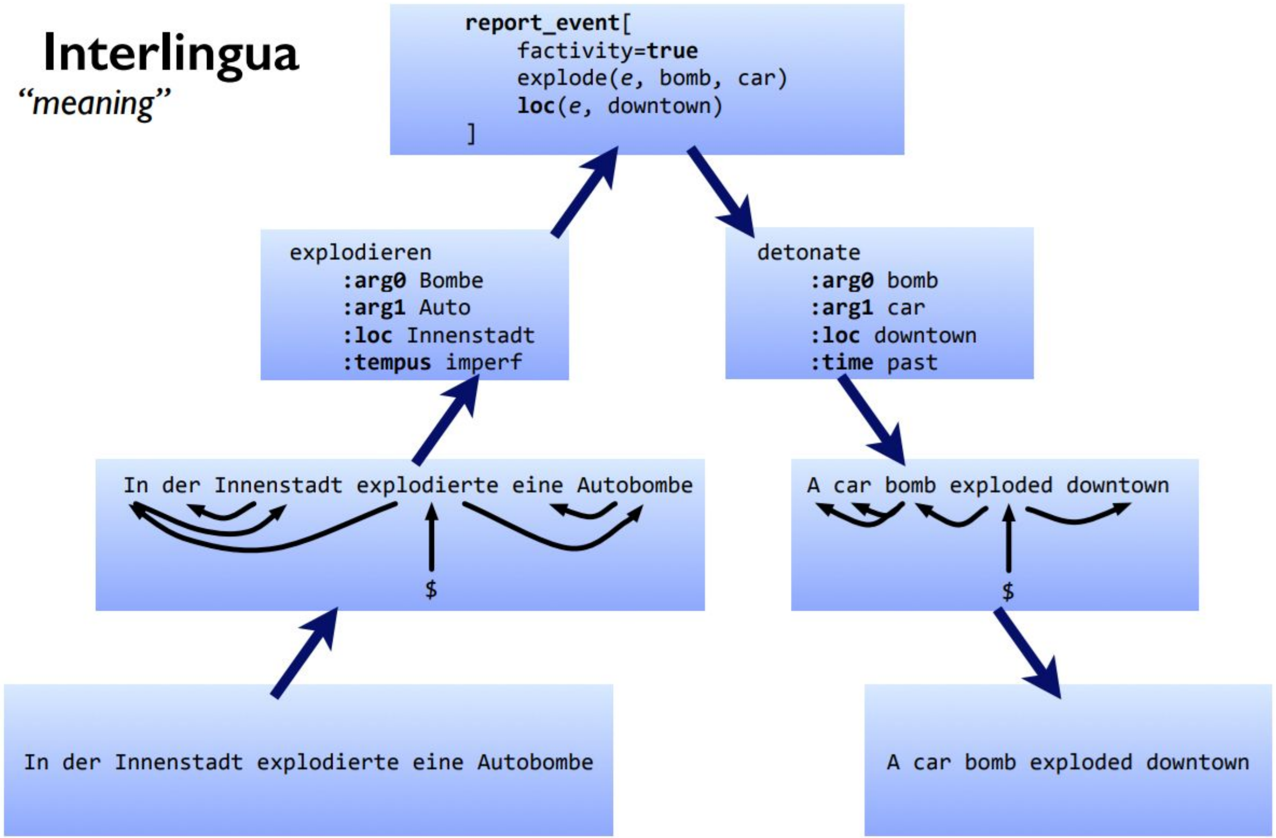
- Semantic transfer

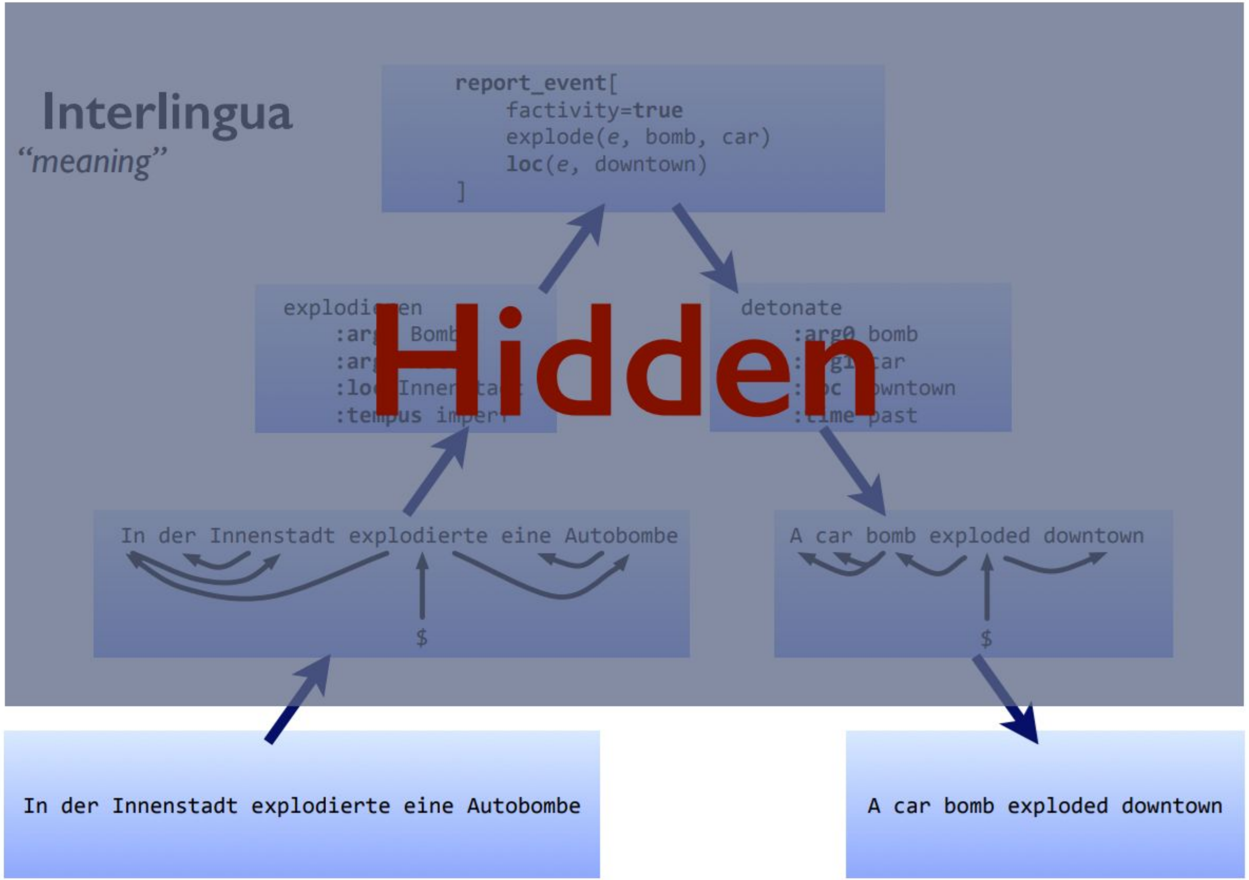


Transfer approaches



Interlingua




Interlingua
"meaning"
report_event[
factivity=true
explode(e, bomb, car)
loc(e, downtown)
]
Hidden
detonate
:arg0 bomb
:time past
In der Innenstadt explodierte eine Autobombe
A car bomb exploded downtown
$
$
In der Innenstadt explodierte eine Autobombe
A car bomb exploded downtown

Learning from data

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .
3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .
4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .
5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .
6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .
7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .
8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .
9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .

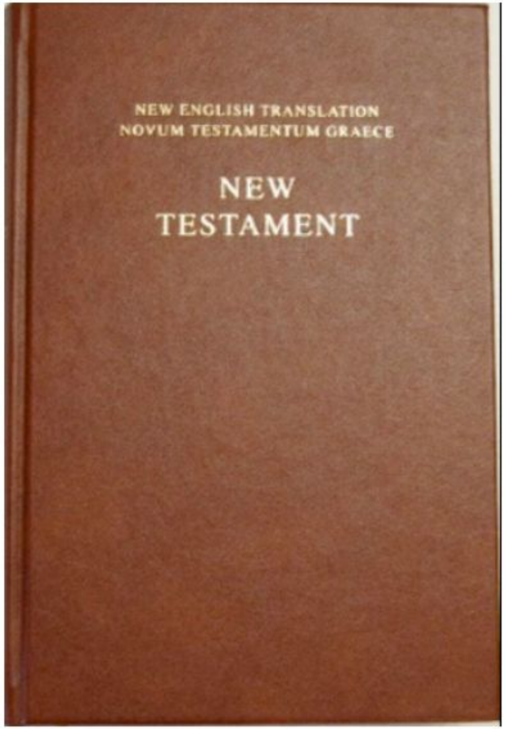
12a. lalok rarok nok izok hihok mok .
12b. wat nnat forat arrat vat gat .

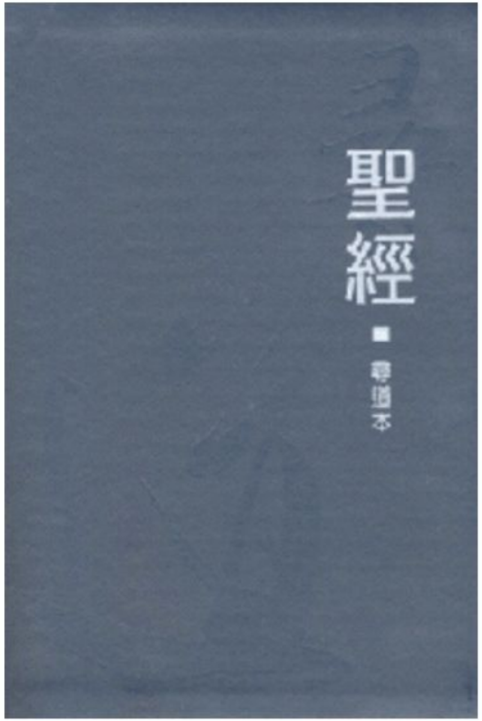
Translation challenge: **farok crrrok hihok yorok clok kantok ok-yurp**

(from Knight (1997): Automating Knowledge Acquisition for Machine Translation)

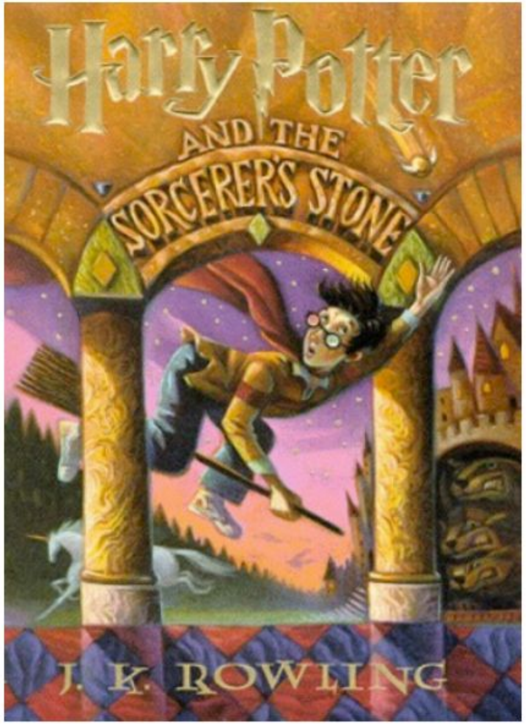
Parallel corpora

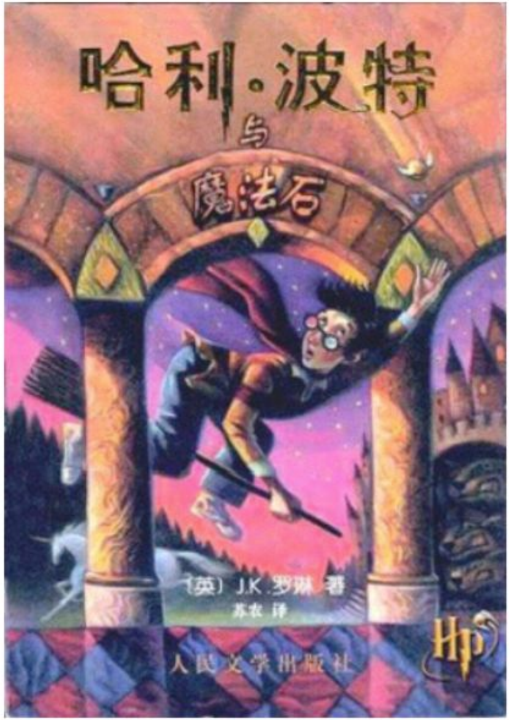
Parallel corpora





Parallel corpora





Parallel corpora

CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
蛋花湯	64.	Egg Drop Soup	1.10	2.10
雲蛋湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

Parallel corpora

	ENGLISH	MANDARIN
1	i **wanna** live in a wes anderson world	我想要生活在Wes Anderson的世界里
2	Chicken soup, corn never truly digests. **TMI**.	鸡汤吧，玉米神马的从来没有真正消化过.恶心
3	To DanielVeuleman **yea iknw imma** work on that	对DanielVeuleman说，是的我知道，我正在向那方面努力
4	**msg 4** Warren G his **cday** is today 1 **yr** older.	发信息给Warren G，今天是他的生日，又老了一岁了。
5	Where **the hell** have you been all these years?	这些年你**TMD**到哪去了
	ENGLISH	ARABIC
6	It's **gonna** be a warm week!	الاسبو ع **الياي** حر
7	onni this gift only **4 u**	أوني هذة الهدية فقط لك
8	sunset in aqaba **:)**	غروب الشمس في العقبة:)
9	RT @MARYAMALKHAWAJA: there is a call for widespread protests in #bahrain **tmrw**	هناك نداء لظاهرات في عدة مناطق غدا

Table 2: Examples of English-Mandarin and English-Arabic sentence pairs. The English-Mandarin sentences were extracted from Sina Weibo and the English-Arabic sentences were extracted from Twitter. Some messages have been shorted to fit into the table. Some interesting aspects of these sentence pairs are marked in bold.

Mining parallel data from microblogs Ling et al. 2013

opus.nlpl.eu

OPUS

... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ... Contributions are very welcome! Please contact <jorg.tiedemann@helsinki.fi >

Search & download resources: -- select -- | -- select -- | all

Latest News

- 2018-02-15: New corpora: ParaCrawl, XhosaNavy
- 2017-11-06: New version: OpenSubtitles2018
- 2017-11-01: New server location: http://opus.nlpl.eu
- 2016-01-08: New version: OpenSubtitles2016
- 2015-10-15: New versions of TED2013, NCv9
- 2014-10-24: New: JRC-Acquis
- 2014-10-20: New: NCv9, TED talks, DGT, WMT
- 2014-08-21: New: Ubuntu, GNOME
- 2014-07-30: New: Translated Books
- 2014-07-27: New: DOGC, Tanzil
- 2014-05-07: Parallel coref corpus ParCor

Search & Browse

- OPUS multilingual search interface
- Europarl v7 search interface
- Europarl v3 search interface
- OpenSubtitles 2016 search interface
- EUconst search interface
- Word Alignment Database (old DB)

Tools & Info

- OPUS Wiki
- OPUS API by Yonathan Koren
- Uplug at bitbucket

Some Projects using OPUS

- Let'sMT! - On-line SMT toolkit
- CASMACAT - Computer Aided Translation

Sub-corpora (downloads & infos):

- Books - A collection of translated literature (Books.tar.gz - 535 MB)
- DGT - A collection of EU Translation Memories provided by the JRC
- DOGC - Documents from the Catalan Goverment (DOGC.tar.gz - 2.8 GB)
- ECB - European Central Bank corpus (ECB.tar.gz - 3.0 GB)
- EMEA - European Medicines Agency documents (EMEA.tar.gz - 13.0 GB)
- The EU bookshop corpus (EUbookshop.tar.gz - 42 GB)
- EUconst - The European constitution (EUconst.tar.gz - 82` MB)
- EUROPARL v7 - European Parliament Proceedings (Europarl.tar.gz - 2(GB)
- GNOME - GNOME localization files (GNOME.tar.gz - 9 GB)
- Global Voices - News stories in various languages (GlobalVoices.tar.gz - 1.2 GB)
- The Croatian - English WaC corpus (hrenWaC.tar.gz - 59 MB)
- JRC-Acquis- legislative EU texts (JRC-Acquis.tar.gz - 11 GB)

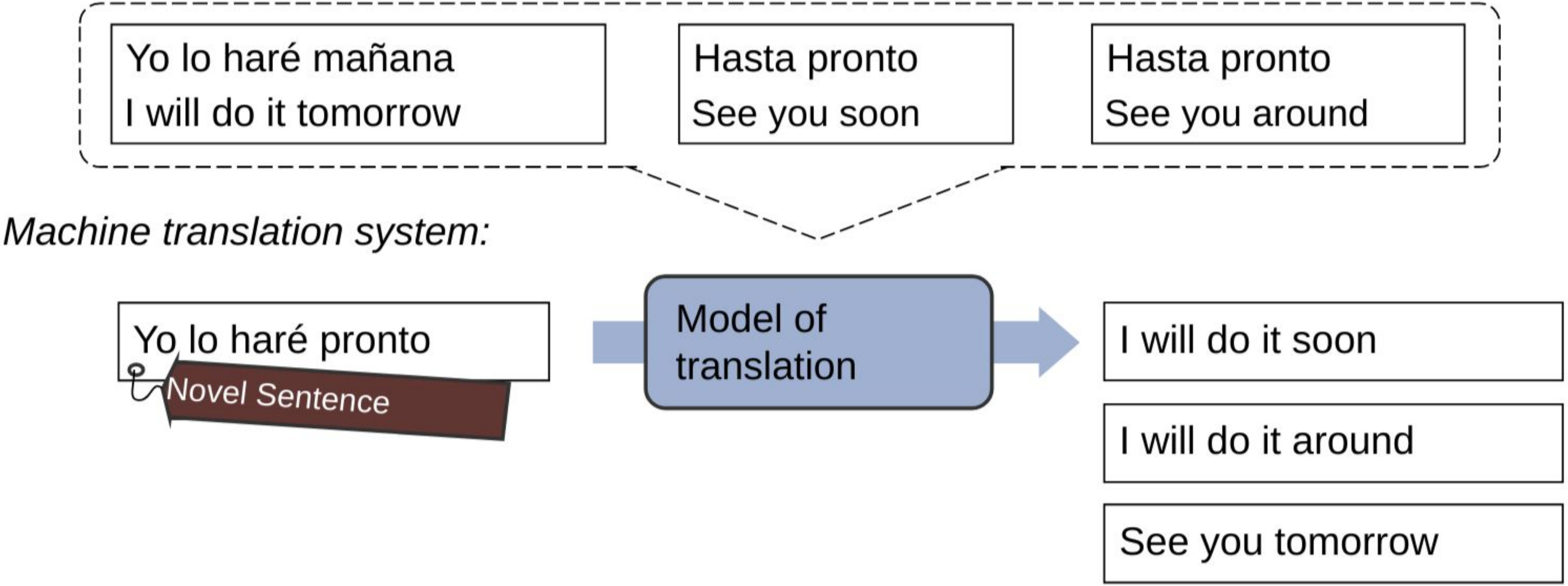
MT as a supervised problem

Sentence-aligned parallel corpus:

Yo lo haré mañana I will do it tomorrow	Hasta pronto See you soon	Hasta pronto See you around

MT as a supervised problem

Sentence-aligned parallel corpus:



Research Problems

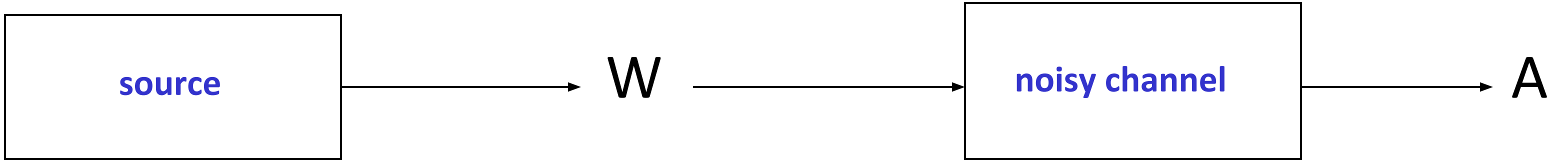
- How can we formalize the process of learning to translate from examples?
- How can we formalize the process of finding translations for new inputs?
- If our model produces many outputs, how do we find the best one?
- If we have a gold standard translation, how can we tell if our output is good or bad?

MT as code breaking

One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘***This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.***’

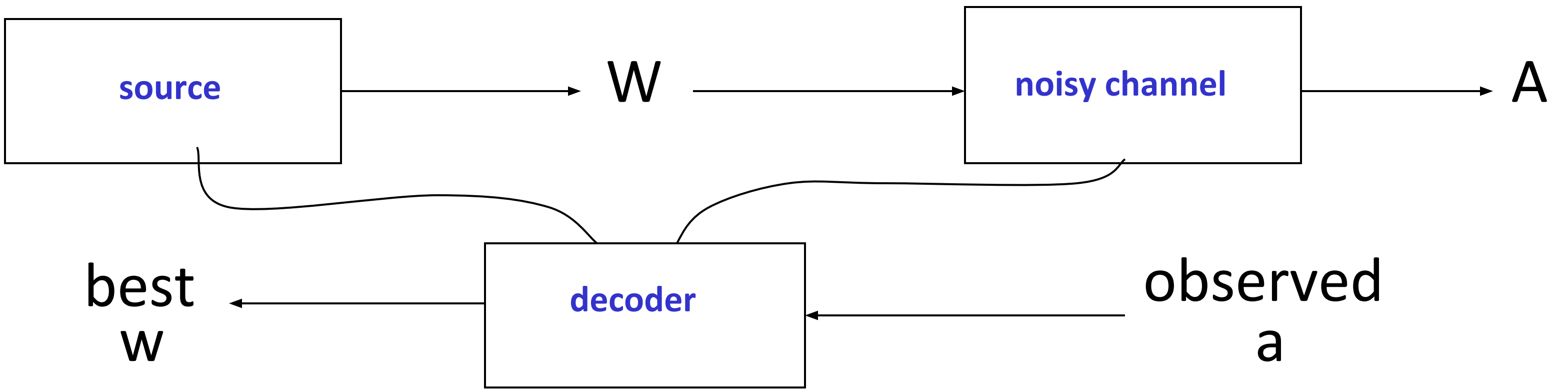
Warren Weaver to Norbert Wiener, March, 1947

The Noisy-Channel Model

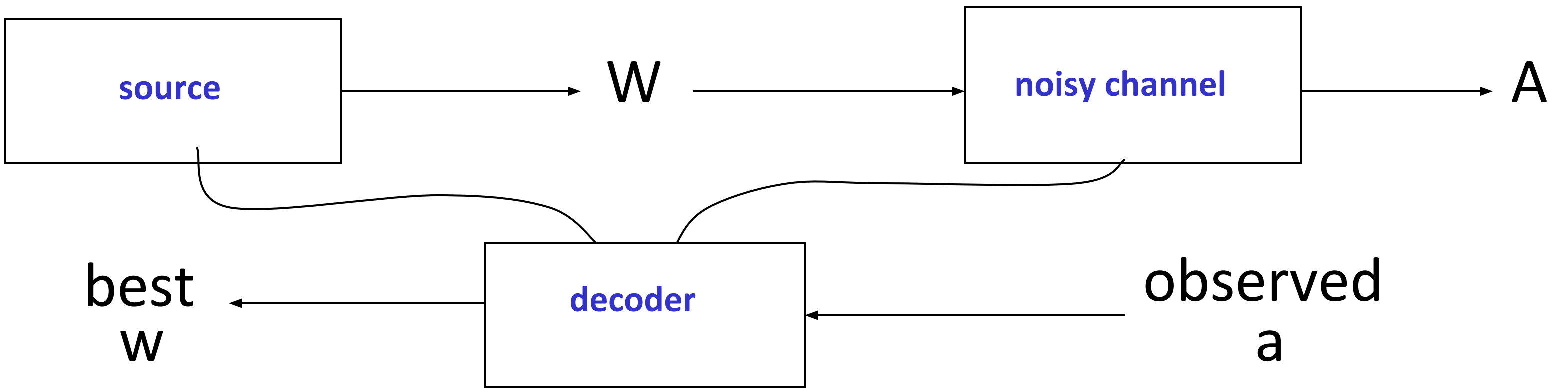


Claude Shannon. "A Mathematical Theory of Communication" 1948.

The Noisy-Channel Model



The Noisy-Channel Model



- We want to predict a sentence given acoustics/foreign language:

$$w^* = \arg\max_w P(w|a)$$

The Noisy-Channel Model

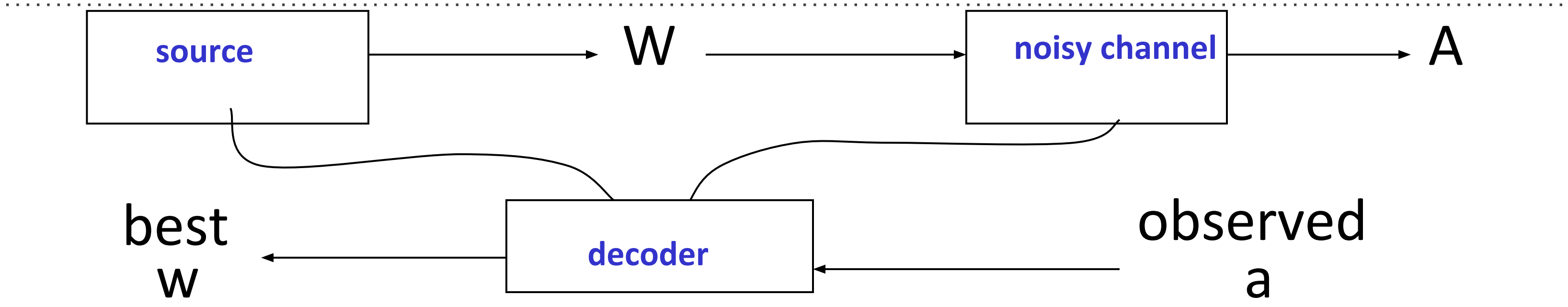
- We want to predict a sentence given acoustics:

$$w^* = \arg\max_w P(w|a)$$

- The noisy-channel approach:

$$\begin{aligned} w^* &= \arg\max_w P(w|a) \\ &= \arg\max_w P(a|w)P(w)/P(a) \end{aligned}$$

The Noisy-Channel Model



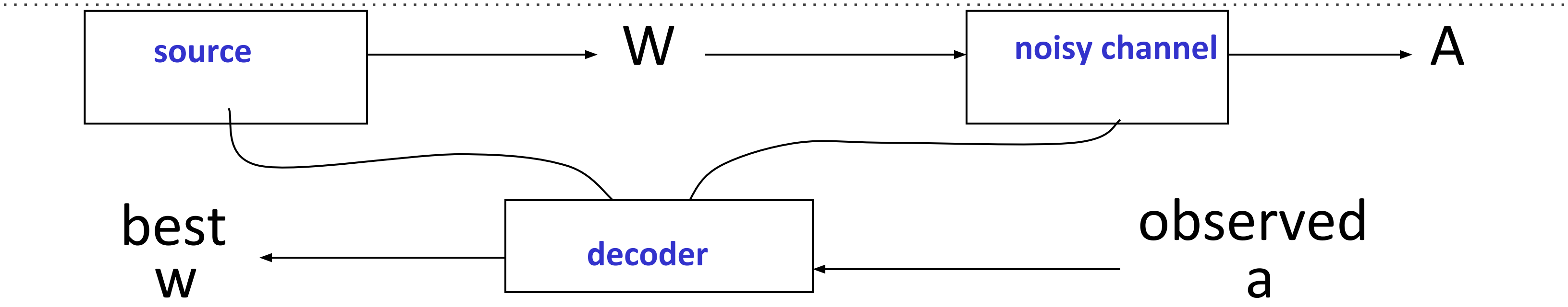
- The noisy-channel approach:

$$
\begin{aligned}
w^* &= \arg\max_w P(w|a) \\
&= \arg\max_w P(a|w)P(w)/P(a) \\
&= \arg\max_w \underbrace{P(a|w)}_{\text{channel model}}\underbrace{P(w)}_{\text{source model}}
\end{aligned}
$$

channel model

source model

The Noisy-Channel Model



- The noisy-channel approach:

$$
\begin{aligned}
w^* &= \arg\max_w P(w|a) \\
&= \arg\max_w P(a|w)P(w)/P(a) \\
&= \arg\max_w P(a|w)P(w)
\end{aligned}
$$

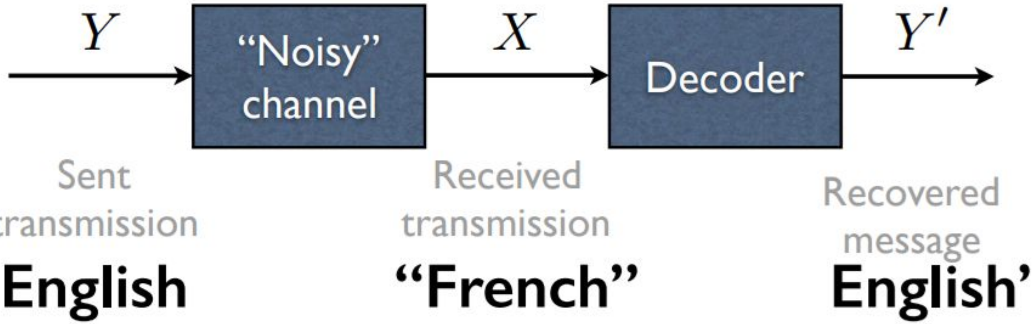
Likelihood

Prior

Acoustic model (HMMs)
Translation model

Language model: Distributions over sequences of words (sentences)

Noisy Channel Model

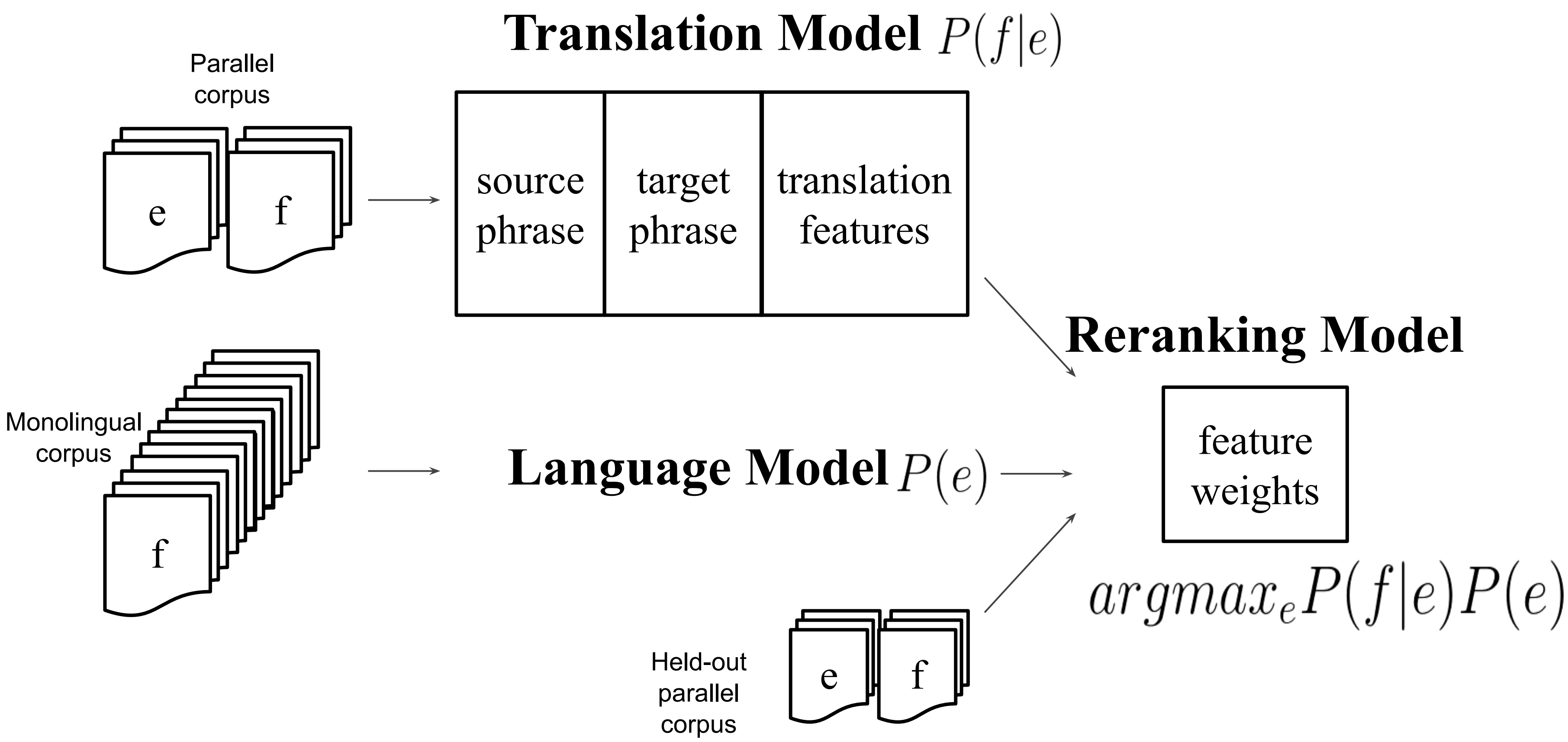


$$\hat{e} = \arg\max_{\boldsymbol{e}} p_{\boldsymbol{\varphi}}(\boldsymbol{e}) \times p_{\boldsymbol{\theta}}(\boldsymbol{f} \mid \boldsymbol{e})$$

language model

translation model

Noisy Channel: Phrase-Based MT



Lexical Translation

в	этом	смысле	подобные	действия	частично	дискредитируют	систему	американской	демократии
in	this	sense	such	actions	some	discredit	system	american	democracy

IBM Model 1	lexical translation
IBM Model 2	adds absolute reordering model
IBM Model 3	adds fertility model
IBM Model 4	relative reordering model
IBM Model 5	fixes deficiency

Phrase-Based Translation

в	этом	смысле	подобные	действия	частично	дискредитируют	систему	американской	демократии
in	this	sense	such	actions	some	discredit	system	american	democracy
the	that	meaning	similar	action	partially		a system	u.s.	democracies
a	the	terms	these	the	part		systems	us	democratic
at	it	way	this	acts	in part		which	america	of democracy
	here	sense ,	like	steps	partly		network	america's	
this			these actions					american democracy	
in this sense								america’s democracy	
in that sense								us democracy	
in this respect									

MT as Direct Modeling

$$\hat{e} = \arg\max_{\boldsymbol{e}} p_{\boldsymbol{\lambda}}(\underbrace{\boldsymbol{e}}_{\text{target}} \mid \underbrace{\boldsymbol{f}}_{\text{source}})$$

- one model does everything
- trained to reproduce a corpus of translations

Conditional Language Modeling

- Calculating the probability of a sentence

$$P(X) = \prod_{i=1}^{I} P(\underbrace{x_i}_{\text{Next Word}} \mid \underbrace{x_1, \ldots, x_{i-1}}_{\text{Context}})$$

Conditional Language Modeling

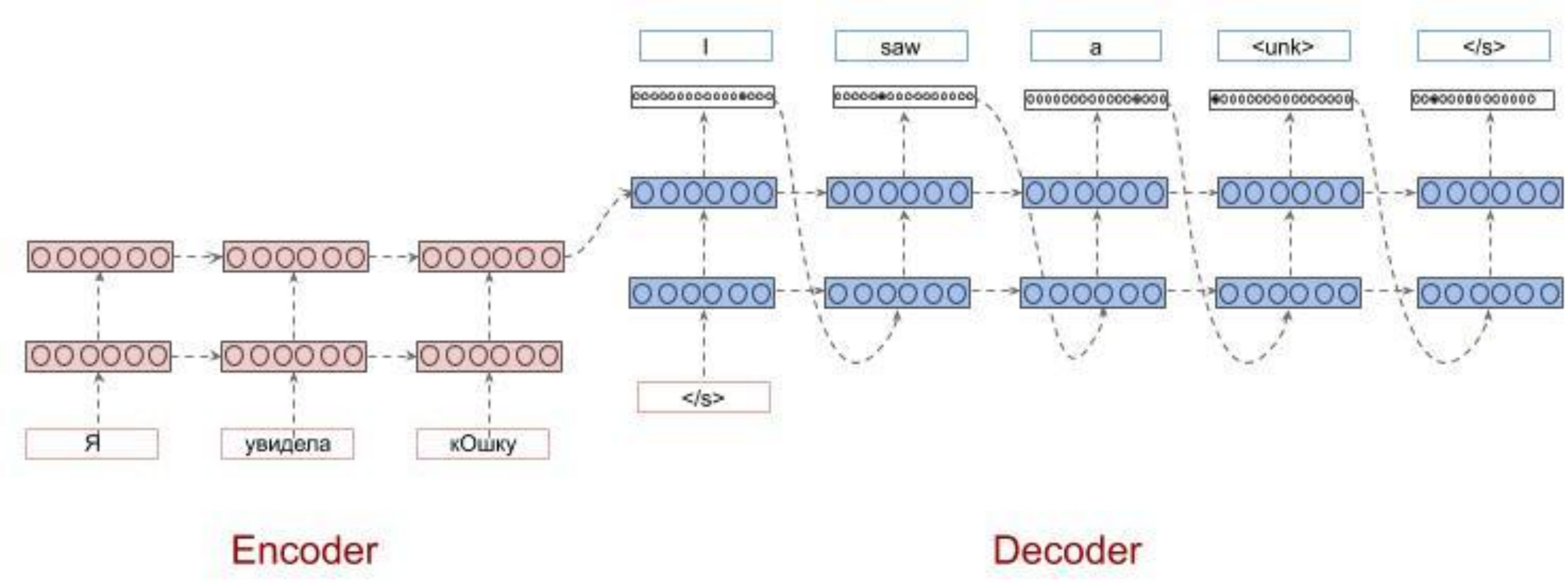
- Calculating the probability of a sentence

$$P(X) = \prod_{i=1}^{I} P(\underbrace{x_i}_{\text{Next Word}} \mid \underbrace{x_1, \ldots, x_{i-1}}_{\text{Context}})$$

- Conditional language models

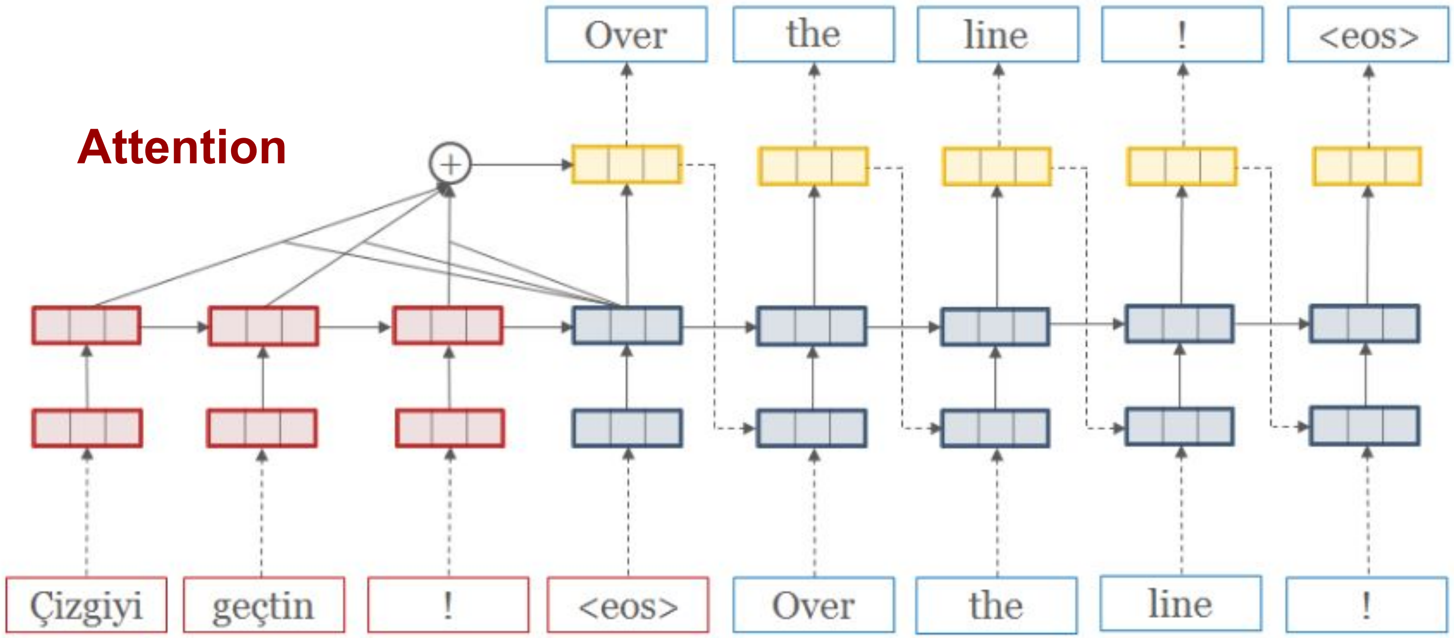
$$P(Y|X) = \prod_{j=1}^{J} P(y_j \mid \underbrace{X}_{\text{Added Context!}}, y_1, \ldots, y_{j-1})$$

Example of neural MT as conditional language model



Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." NeurIPS 2014.

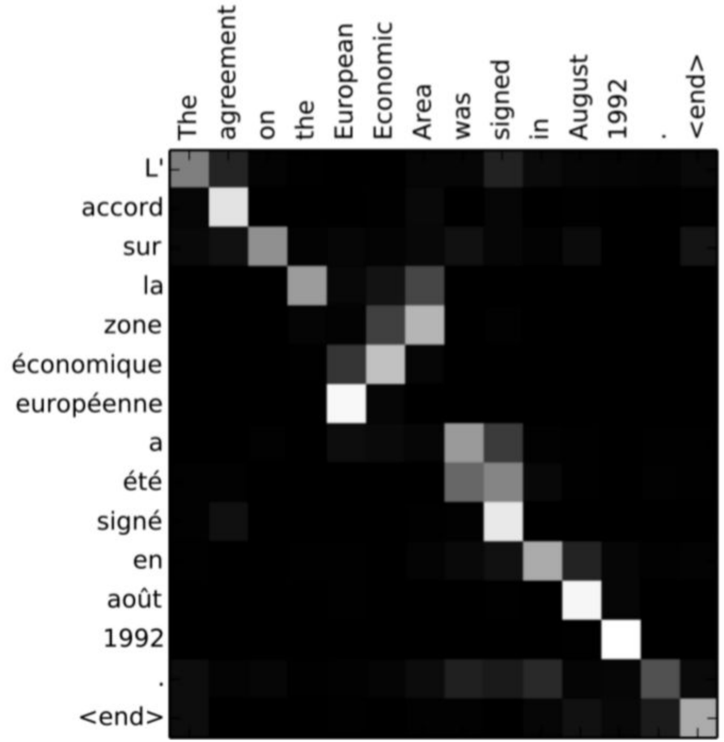
Attention mechanism



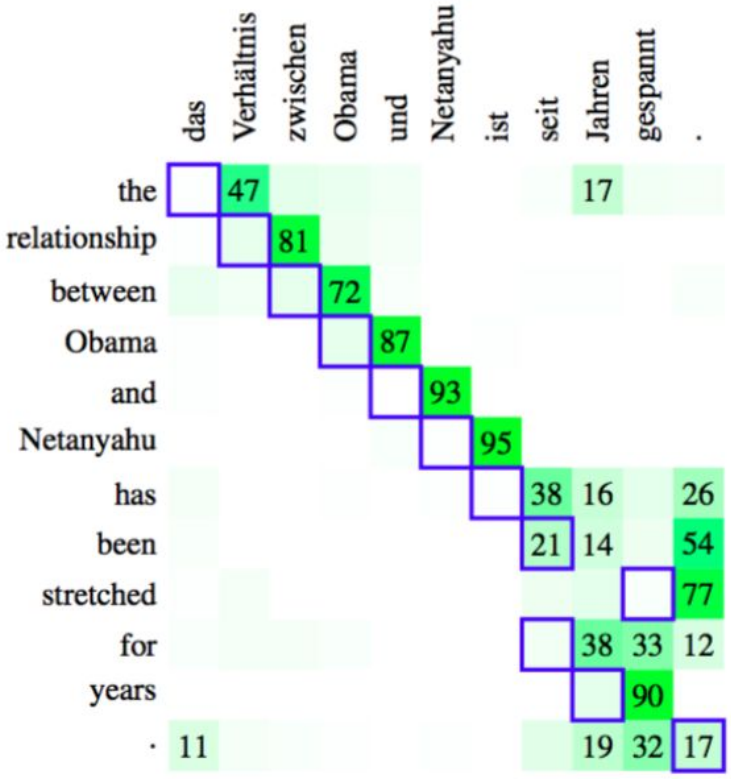
http://opennmt.net/

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." ICLR 2015

Attention Bahdanau et al. (2015)

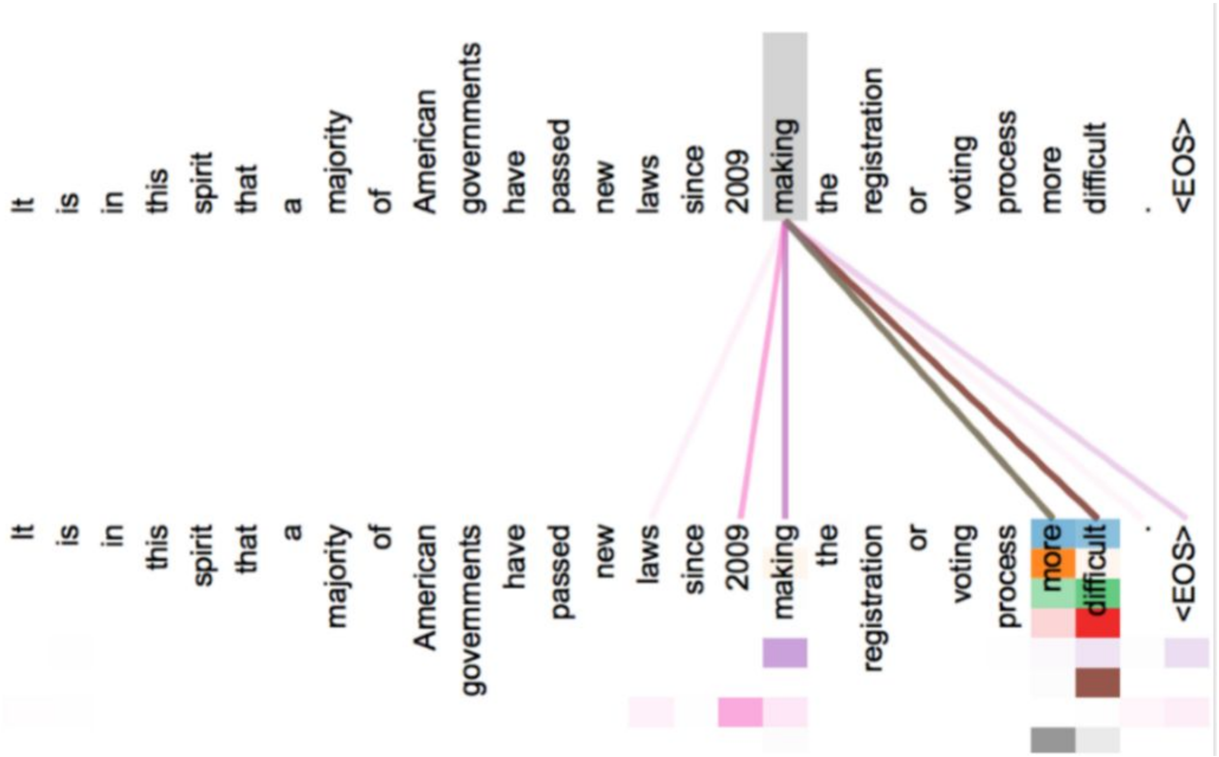


Attention is not alignment



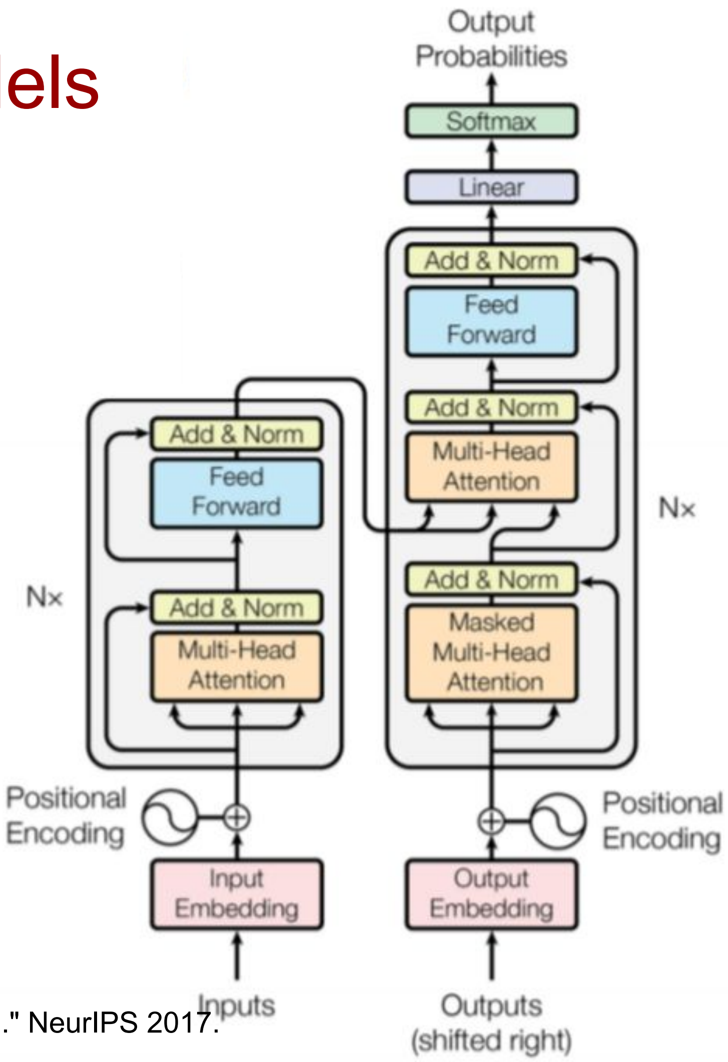
Philipp Koehn and Rebecca Knowles. "Six challenges for neural machine translation." WNGT 2017

Multi-headed attention



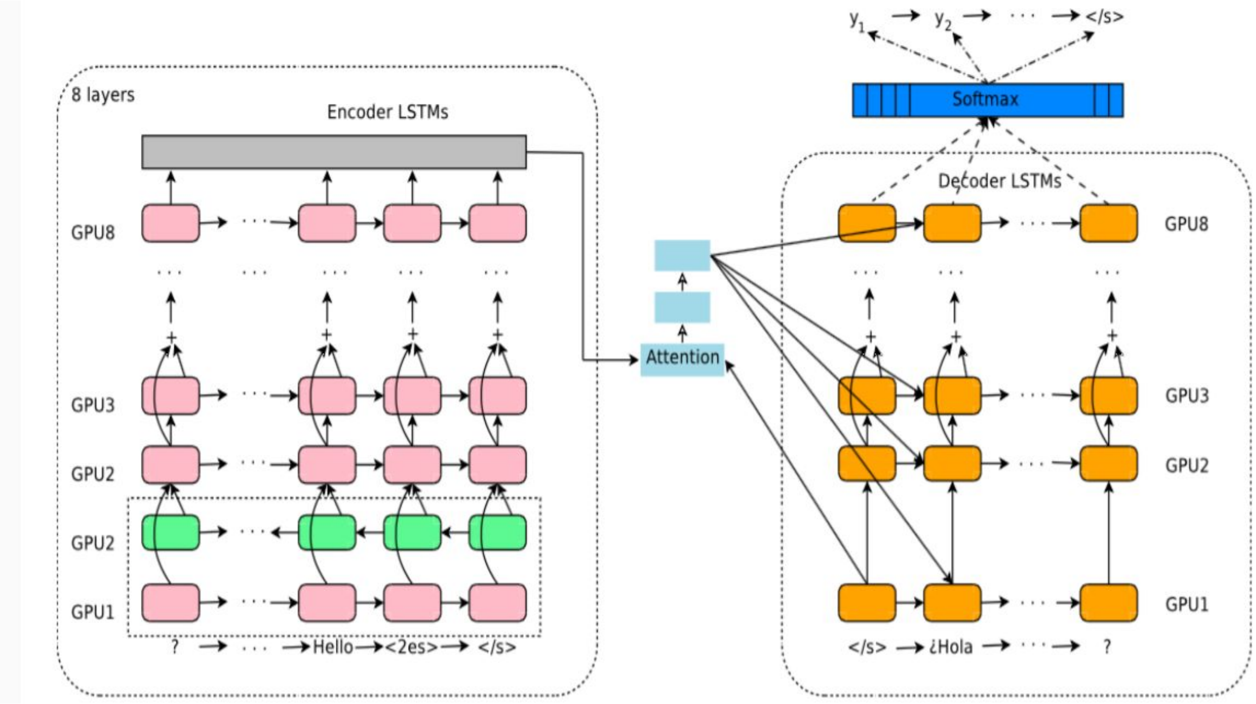
Vaswani, Ashish, et al. "Attention is all you need." NeurIPS 2017.

Transformer models



Vaswani, Ashish, et al. "Attention is all you need." NeurIPS 2017.

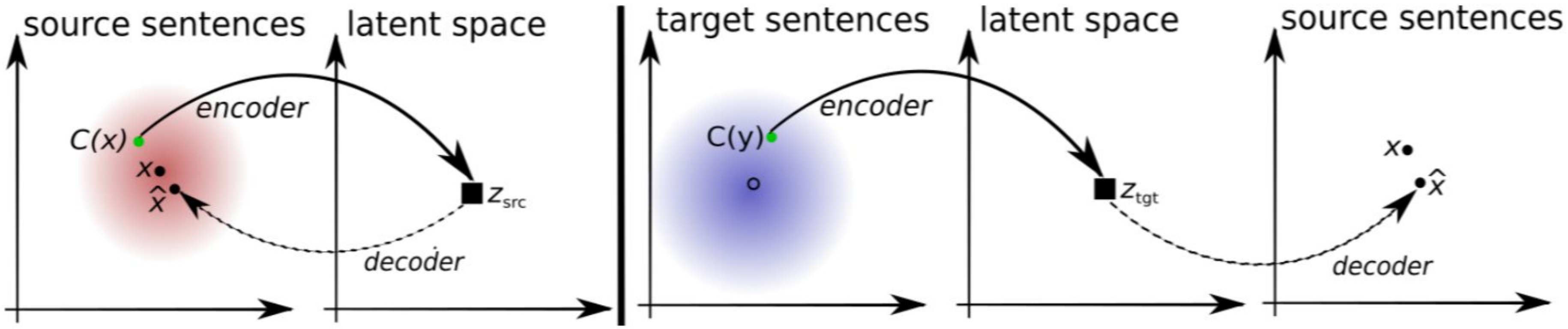
Multilingual MT



$\boxed{\text{<2es>}}$ Hello, how are you? -> ¿Hola como estás?

Johnson, Melvin, et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation."TACL 2017.

Unsupervised MT



Denoising autoencoder

Back-translation:

- translate target to source
- use as a “supervised” example to translate source to target

Guillaume Lample, et al. "Unsupervised Machine Translation Using Monolingual Corpora Only." ICLR 2018.

Two Views of MT

- Code breaking (aka the noisy channel, Bayes rule)
  - I know the target language
  - I have example translations texts (example enciphered data)

- Direct modeling (aka pattern matching)
  - I have really good learning algorithms and a bunch of example inputs (source language sentences) and outputs (target language translations)

Two Views of MT

$$\hat{\mathbf{e}} = \arg\max_{\boldsymbol{e}} p_{\boldsymbol{\varphi}}(\boldsymbol{e}) \times p_{\boldsymbol{\theta}}(\boldsymbol{f} \mid \boldsymbol{e})$$ Noisy channel

$$\hat{\mathbf{e}} = \arg\max_{\boldsymbol{e}} p_{\boldsymbol{\lambda}}(\boldsymbol{e} \mid \boldsymbol{f})$$ Direct

A common problem

$$\hat{e} = \arg\max_{e} p_{\varphi}(e) \times p_{\boldsymbol{\theta}}(\boldsymbol{f} \mid \boldsymbol{e})$$ Noisy channel

$$\hat{e} = \arg\max_{e} p_{\boldsymbol{\lambda}}(\boldsymbol{e} \mid \boldsymbol{f})$$ Direct

Both models must assign probabilities to how a sentence in one language translates into a sentence in another language.

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} p_{\boldsymbol{\varphi}}(\boldsymbol{e}) \times p_{\boldsymbol{\theta}}(\boldsymbol{f} \mid \boldsymbol{e})$$ Noisy channel

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} p_{\boldsymbol{\lambda}}(\boldsymbol{e} \mid \boldsymbol{f})$$ Direct

Which is better?

- Noisy channel - $p_{\boldsymbol{\theta}}(\boldsymbol{e}) \times p_{\boldsymbol{\varphi}}(\boldsymbol{f} \mid \boldsymbol{e})$
  - easy to use monolingual target language data
  - search happens under a product of two models (individual models can be simple, product can be powerful)
  - obtaining probabilities requires renormalizing

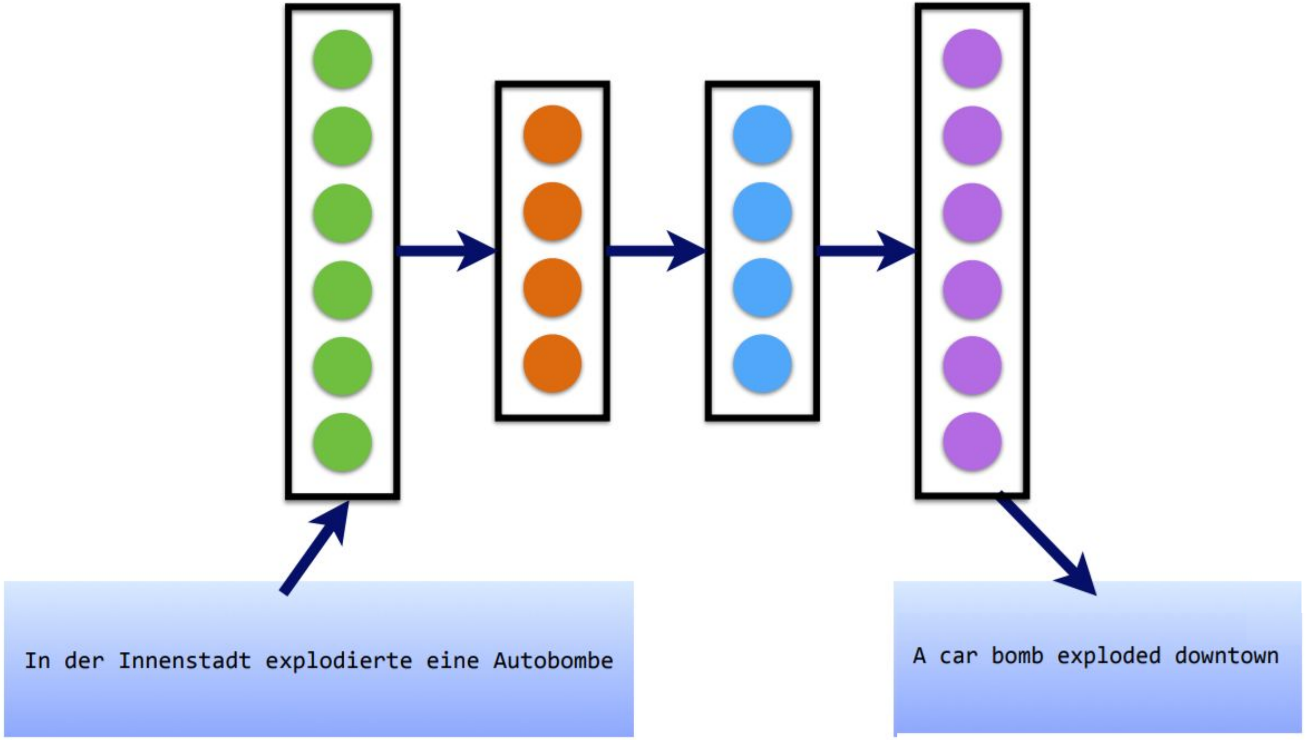
- Direct model - $p_{\boldsymbol{\lambda}}(\boldsymbol{e} \mid \boldsymbol{f})$
  - directly model the process you care about
  - model must be very powerful

Where are we in 2020?

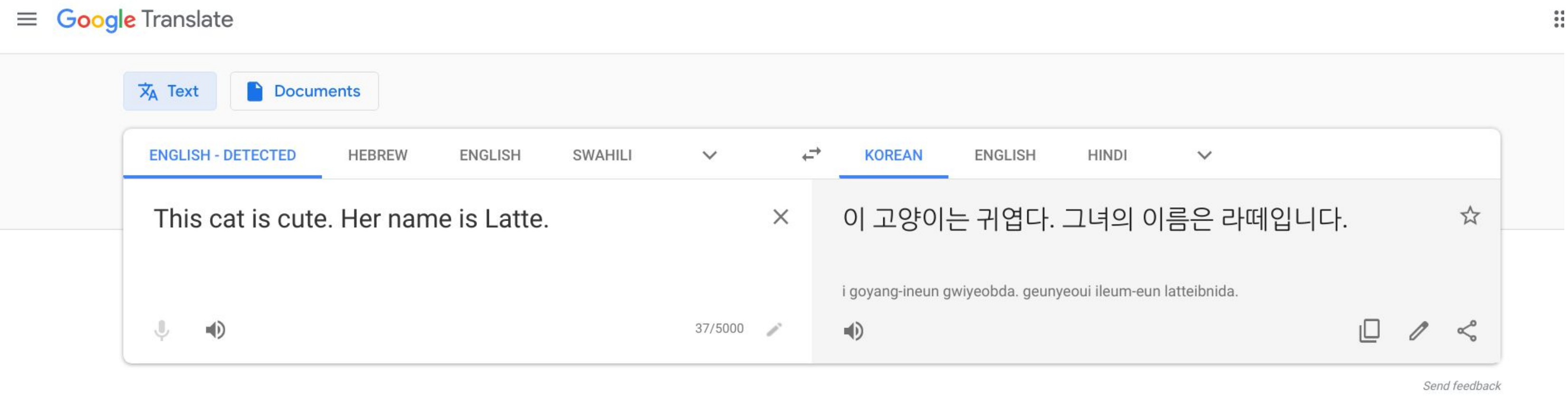
- Direct modeling is where most of the action is
  - Neural networks are very good at generalizing and conceptually very simple
  - Inference in “product of two models” is hard

- Noisy channel ideas are incredibly important as they play a role in how we think about translation

Interlingua?



Is it a good translation?



MT evaluation is hard

- MT Evaluation is a research topic on its own
- Language variability: there is no single correct translation
  - Is system A better than system B?
- Human evaluation is subjective

Human evaluation

- Adequacy and Fluency
  - Usually on a Likert scale (1 “not adequate at all” to 5 “completely adequate”)

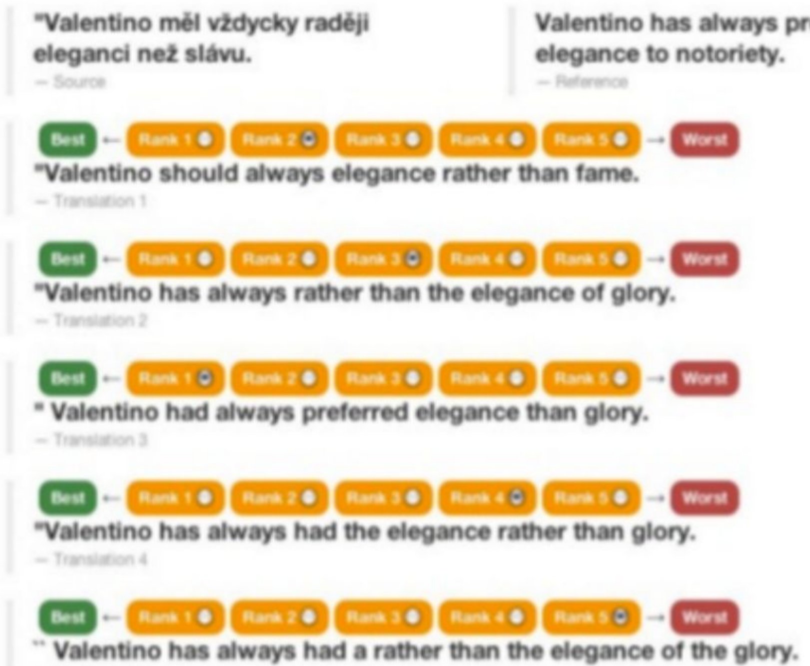
Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

Human evaluation

- Ranking of the outputs of different systems at the system level

WMT-13 Appraise tool: rank translations best-worst (w. ties)



Human evaluation

- Adequacy and Fluency
  - Usually on a Likert scale (1 “not adequate at all” to 5 “completely adequate”)
- Ranking of the outputs of different systems at the system level
- Post editing effort: how much effort does it take for a translator (or even monolingual) to “fix” the MT output so it is “good”
- Task-based evaluation: was the performance of the MT system sufficient to perform a task.

Automatic evaluation

- Precision-based
  - **BLEU**, NIST, ...
- F-score-based
  - Meteor,...
- Error rates
  - WER, TER, PER,...
- Using syntax/semantics
  - PosBleu, Meant, DepRef,...
- Embedding based
  - BertScore, **chrF, YISI-1, ESIM**, ...

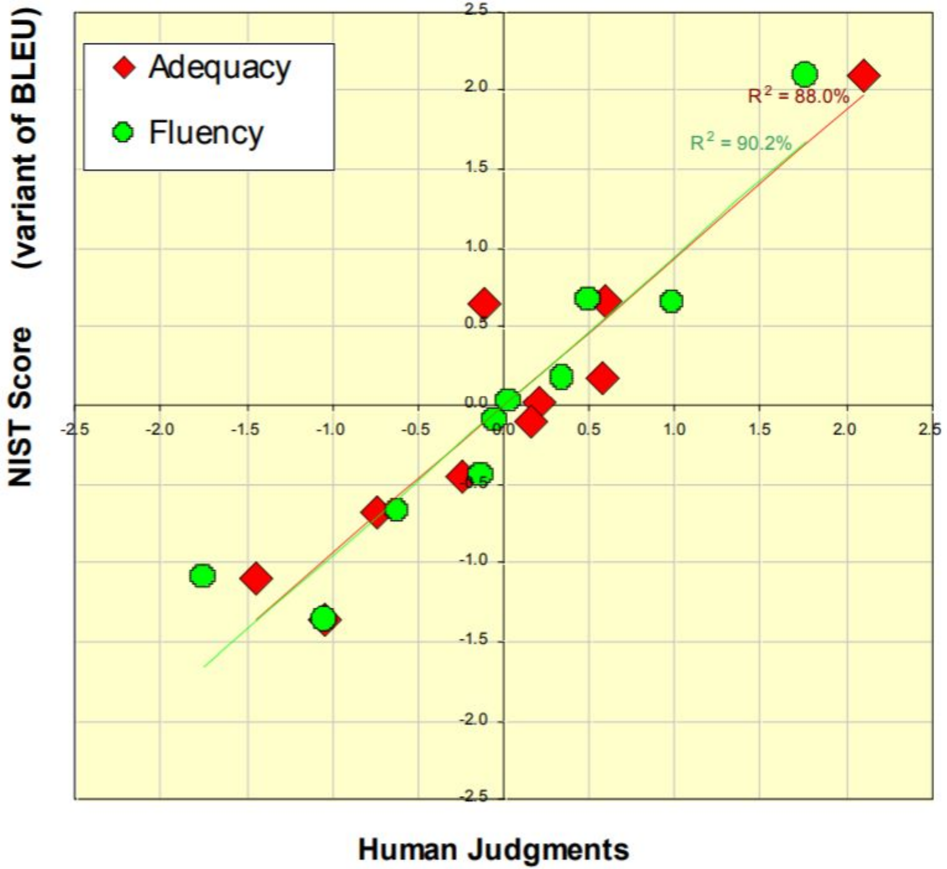
Automatic evaluation

- The BLEU score proposed by IBM (Papineni et al., 2002)
  - Count n-grams overlap between machine translation output and reference reference translations
  - Compute precision for ngrams of size 1 to 4
  - No recall (because difficult with multiple references)
  - To compensate for recall: “brevity penalty”. Translations that are too short are penalized
  - Final score is the geometric average of the n-gram precisions, times the brevity penalty

$$\mathrm{BLEU} = min(1, \frac{output\ length}{reference\ length})(\prod_{i-1}^{4} precision_i)^{\frac{1}{4}}$$

  - Calculate the aggregate score over a large test set

BLEU vs. human judgments



Automatic evaluation

- Embedding based
  - BertScore, **chrF, YISI-1, ESIM**, ...

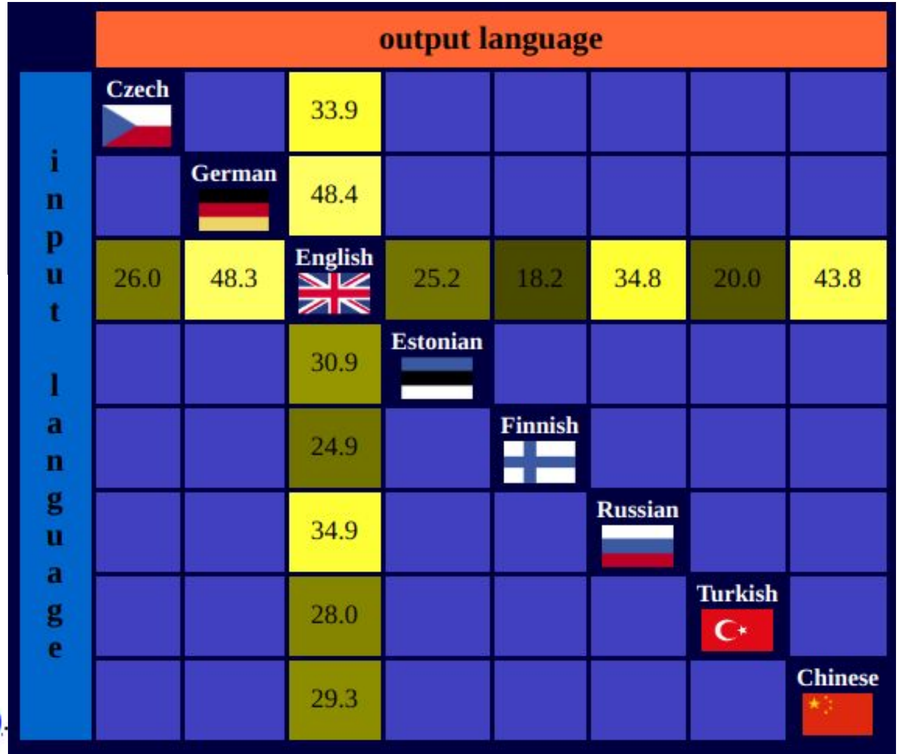
Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

Nitika Mathur **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
nmathur@student.unimelb.edu.au {tbaldwin,tcohn}@unimelb.edu.au

MT venues and competitions

- MT tracks in *CL conferences
- **WMT**, IWSLT, AMTA...

- www.statmt.org



- the NAACL-2006 Workshop on Statistical Machine Translation,
- the ACL-2007 Workshop on Statistical Machine Translation,
- the ACL-2008 Workshop on Statistical Machine Translation,
- the EACL-2009 Workshop on Statistical Machine Translation,
- the ACL-2010 Workshop on Statistical Machine Translation
- the EMNLP-2011 Workshop on Statistical Machine Translation,
- the NAACL-2012 Workshop on Statistical Machine Translation,
- the ACL-2013 Workshop on Statistical Machine Translation,
- the ACL-2014 Workshop on Statistical Machine Translation,
- the EMNLP-2015 Workshop on Statistical Machine Translation,
- the First Conference on Machine Translation (at ACL-2016).
- the Second Conference on Machine Translation (at EMNLP-2017).