CS11-737 Multilingual NLP

Machine Translation Data and Evaluation

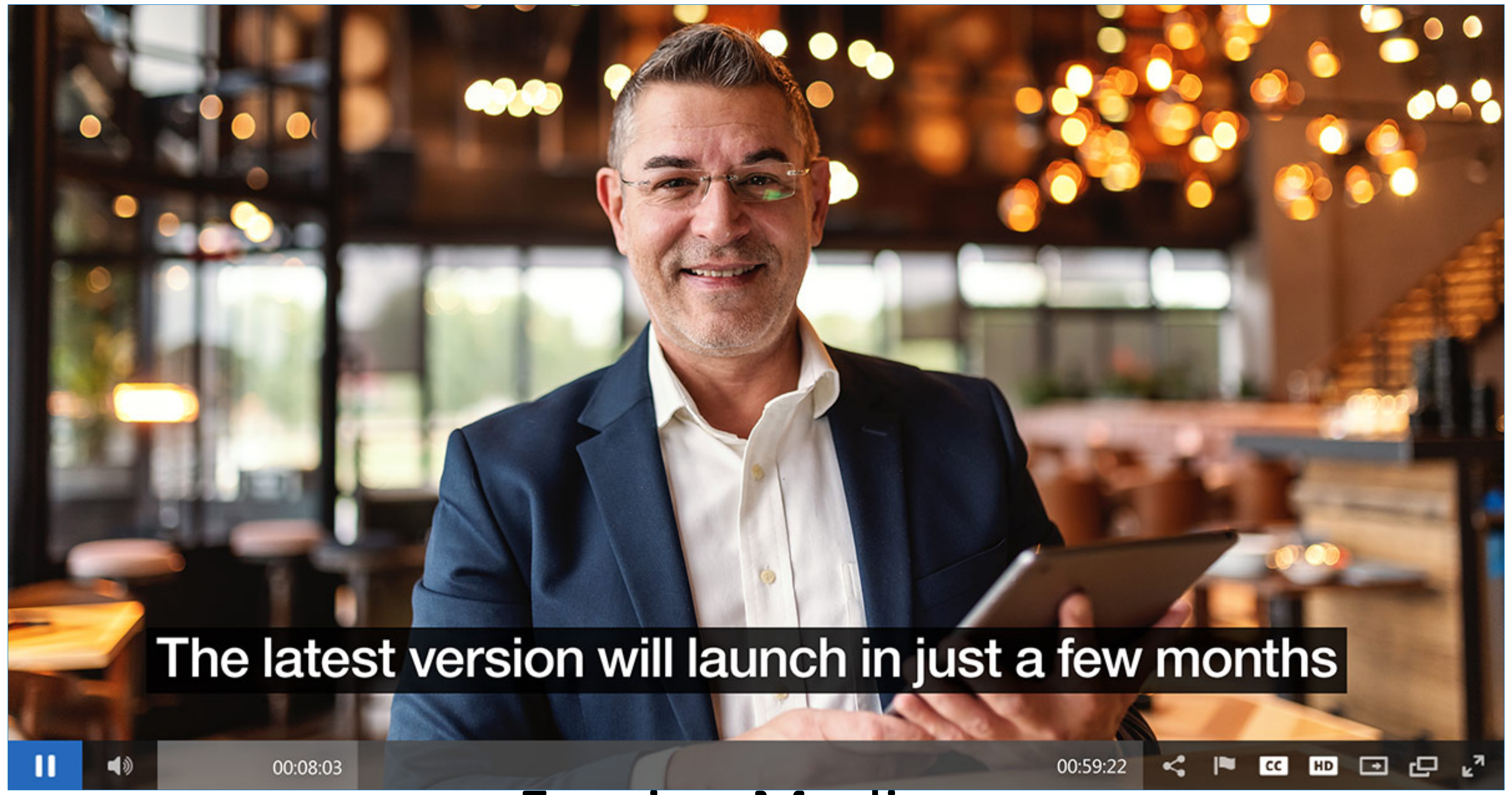
Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

Carnegie Mellon University
Language Technologies Institute

adapted from Yulia Tsvetkov and Graham Neubig

Cross Language Barrier with Machine Translation



Foreign Media

Global Conferences

Tourism

International Trade and e-commerce

Machine Translation has increased international trade by over 10%

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 65, No. 12, December 2019, pp. 5449–5460
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform

Erik Brynjolfsson,[a] **Xiang Hui,**[b] **Meng Liu**[b]

[a] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [b] Marketing, Olin School of Business, Washington University in St. Louis, St. Louis, Missouri 63130

Contact: erikb@mit.edu, http://orcid.org/0000-0002-8031-6990 (EB); hui@wustl.edu, http://orcid.org/0000-0001-7595-3461 (XH); mengl@wustl.edu, http://orcid.org/0000-0002-5512-7952 (ML)

Received: April 18, 2019
Revised: April 18, 2019
Accepted: April 18, 2019
Published Online in Articles in Advance: September 3, 2019

https://doi.org/10.1287/mnsc.2019.3388

Abstract. Artificial intelligence (AI) is surpassing human performance in a growing number of domains. However, there is limited evidence of its economic effects. Using data from a digital platform, we study a key application of AI: machine translation. We find that the introduction of a new machine translation system has significantly increased international trade on this platform, increasing exports by 10.9%. Furthermore, heterogeneous treatment effects are consistent with a substantial reduction in translation costs. Our results provide causal evidence that language barriers significantly hinder trade and that AI has already begun to improve economic efficiency in at least one domain.

History: Accepted by Joshua Gans, business strategy.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2019.3388.

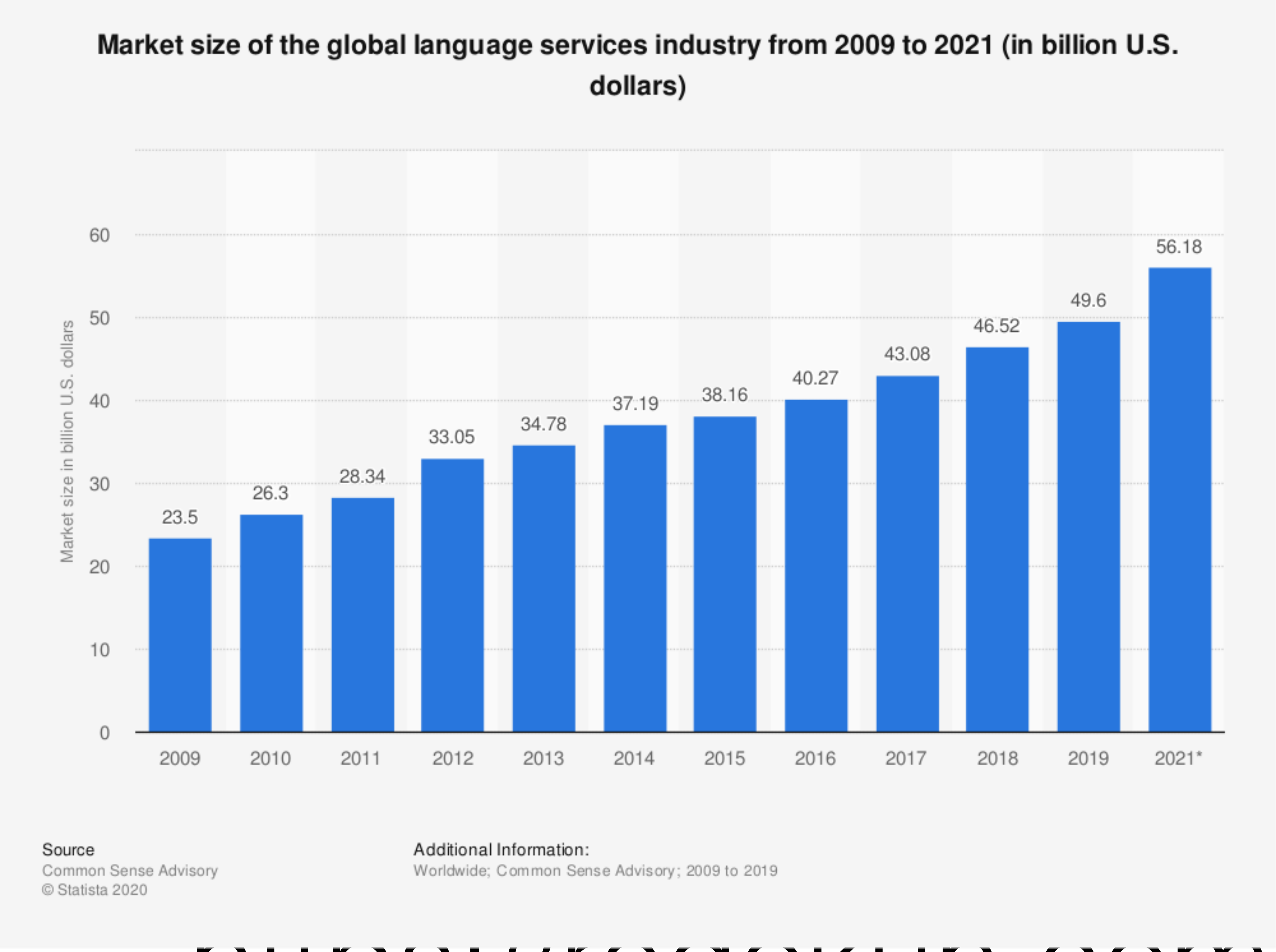
Keywords: artificial intelligence • international trade • machine translation • machine learning • digital platforms

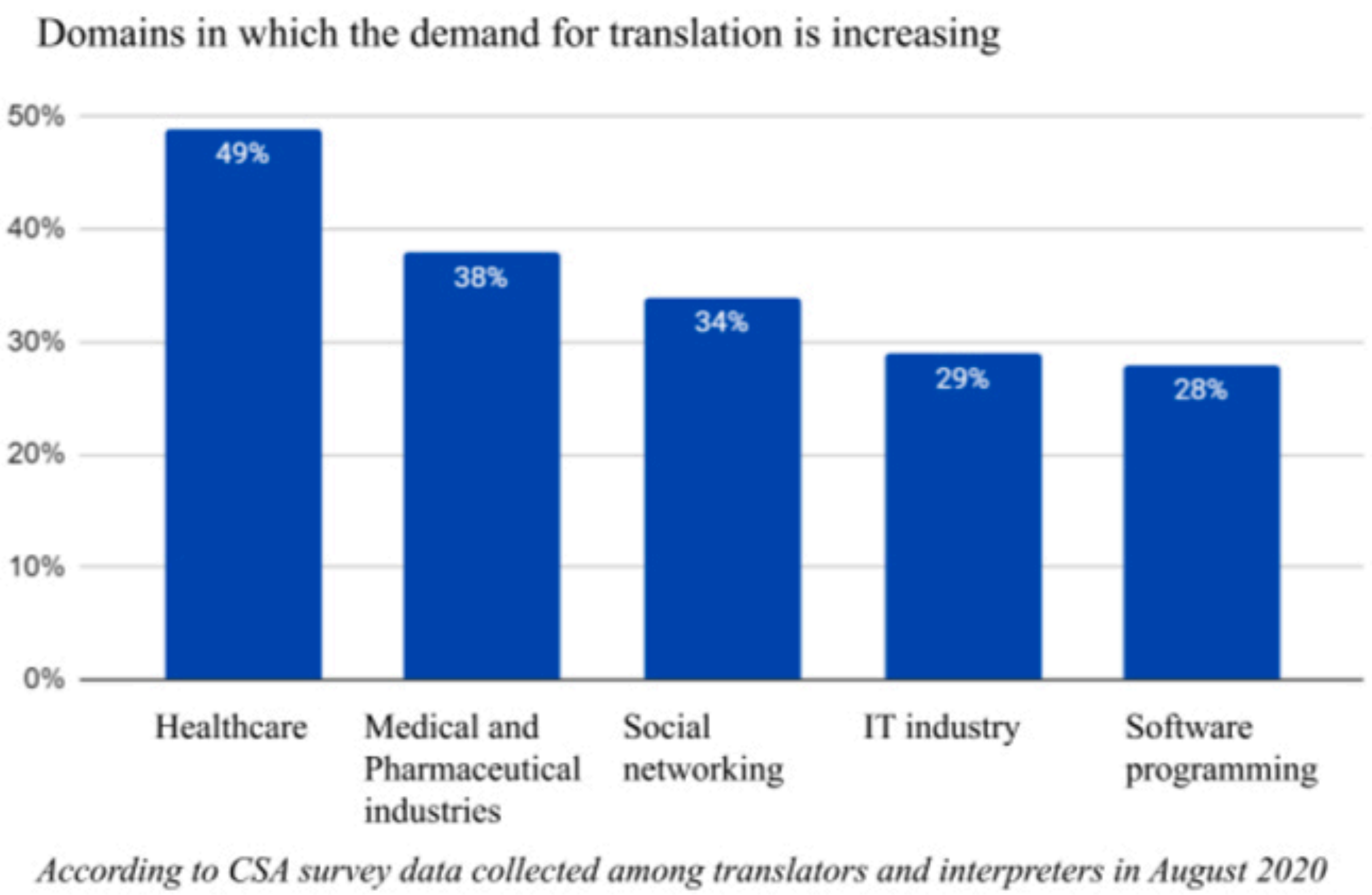
Equivalent to make the world smaller than 26%

study on ebay

Translation Market

- Language services market w/ $60billion (translation, interpretation, MT) in 2022. $9.7 billion US alone.
- 640,000 translators worldwide (about 75% freelance)
- Machine Translation market: $982.2 million





https://redokun.com/blog/translation-statistics

When you really need Machine Translation

- Rimi Natsukawa live streaming on Tiktok July, 2021


INA 0 5
CHN 0 10
TOKYO 2020
TOKYO 2020
BWF
OMEGA
5-10
OMEGA
5-10
TOKYO 2020
TOKYO 2020

Machine Translation

Translating information from one language to another

I bought a sweet persimmon in the store

Ich kaufte eine süße Persimone im laden

Types of Machine Translation

- Translating information from one language to another
- Media:
  - (Text) Machine Translation
  - Speech Translation: Speech-to-Text or Speech-to-speech translation
  - Visually Machine Translation: Text translation with additional image
- Genre:
  - Sentence level MT
  - Document level MT
  - Dialog Translation
- Number of Languages:
  - Bilingual
  - Multilingual

Why automatic Machine Translation?

- Too expensive to hire human translator
    - e.g. touring, shopping, restaurant eating in a foreign country
- Too much effort for human to translate massive text
    - can tolerate imprecise translation
- Need instantaneous translation
    - e.g. in international conference

Outline

- History of Machine Translation
- Challenge of Machine Translation
- Machine translation math framework
- MT data
- MT evaluation

History of MT

A Brief History of Machine Translation

1947

translation as decoding in cryptography

— Warren Weaver

1954

Rule-based MT:

Georgetown-IBM automatic translation of 60 sentences Ru->En

1966

ALPAC report:

MT winter

1968

Systran

1976

METEO system for weather forecasts in Canada

En->Fr

1984

Example-based MT

Makoko Nagao

1980s - 2000s

Statistical MT (SMT)

Moses, Google

2014, 2015, 2017

Neural MT (NMT)

Seq2Seq

Attention

Transformer

History of Machine Translation

- Warren Weaver: translation as 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."
(1947, in a letter to Norbert Wiener)

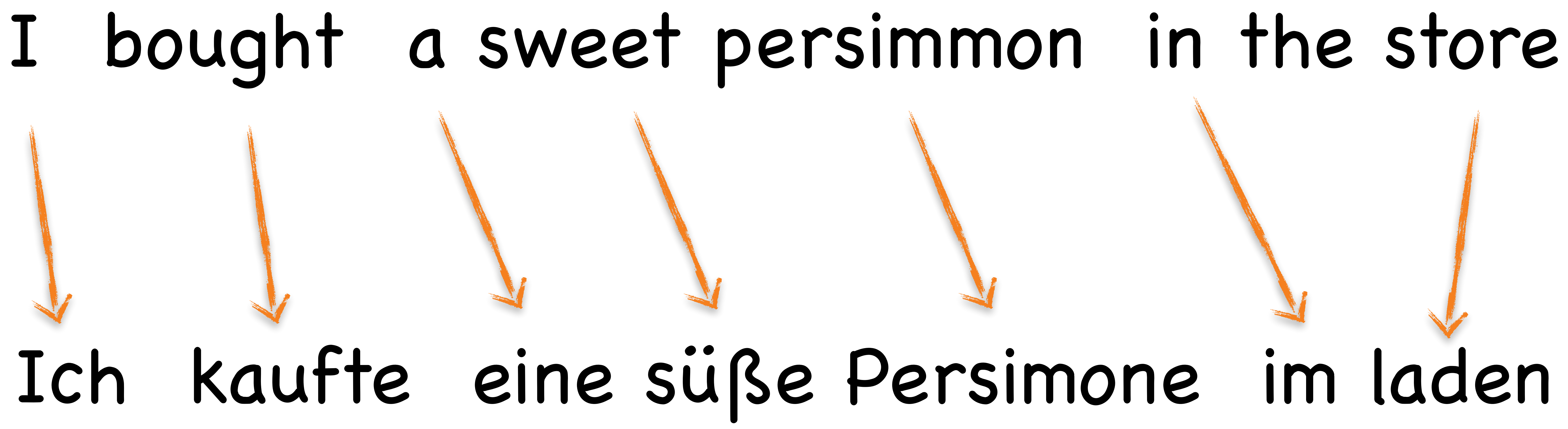
1950s-1960s

- 1954: Georgetown-IBM experiment, aut
  of 60 Russian sentences into English, u
  - Only 6 grammar rules and 250 tokens.
  - W. John Hutchins , Leon Dostert , Paul Garvi

- 1966 ALPAC report
  - We do not have useful machine translation ar
    immediate or predictable prospect of useful machine translation
  - Funding cut for MT in US in the following 20 yrs

Rule-based System

- METEO system for weather forecasts (1976)
  - Used by Environment Canada from 1981 to 2001, to translate between English and French
- Systran (1968)



Example-based Machine Translation

- 1984: Makoto Nagao, A framework of mechanical translation between Japanese and English by analogy principle

How much is that **red umbrella** ? = Ano **akai kasa** wa ikura desu ka.

How much is that small camera ? ? Ano chiisai kamera wa ikura desu ka.

Statistical Machine Translation

- late 1980s-1990s: IBM
- 2000s: phrase-based MT (Moses, Google)
- Training statistical model from parallel corpus

$$\operatorname{argmax} p_\theta(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

$p(x \mid y)$: translation model, $p(y)$: language model

I bought a sweet persimmon in the store

Ich kaufte eine süße Persimone im laden

Neural Machine Translation

- Trained in end-to-end fashion (no intermediate separate training)
- 2014: Sequence to sequence learning with neural networks
  - Define LSTM encoder-decoder framework
- 2015: Neural Machine Translation by Jointly Learning to Align and Translate
  - Define attention mechanism between encoder-decoder
- 2016: Google translate deploys NMT
- 2017: Attention is all you need
  - Replace LSTM with multihead attention layers (Transformer)
- Almost all major production MT systems use NMT now

Commercial Machine Translation

- Google translate: 133 languages, separate app, support text/document translation, image translation, and speech translation
- Microsoft translate: 129 languages for text
- Baidu translate: 200+ languages
- ByteDance VolcTrans: 122 languages
- DeepL: good at European languages, 31 languages
- Youdao Translate: integrated with its own dictionary app
- Tencent Translate: native in wechat, and separate app
- NiuTrans: specialized in Chinese to many languages
- ChatGPT

MT Products for users/clients

Product	User	Scenario	Advantage
Web translate tool, Translation function on Youtube/Tiktok/Twitter/ Facebook	consumers/users who do not know the source language	could tolerate imprecision	convenient, free/low-cost
Computer Aided Translation tools	content creator, translators, knowing both languages	need high precision	productivity and efficiency, additional functionality like translation memory, glossary
translation API e.g. Amazon translation	business client	cost/effective	robust api, easy to integrate and maintain
private MT deployment e.g. NiuTrans	business client		domain-specific models, tailored to special needs
Special MT hardware, e.g. translation pen Simultaneous translation earphone	consumers for targeted scenario		

MT is not just about Model

- User-oriented Product
  - What are real users' needs?
  - Observe how the users are using our product, e.g. how translators are using CAT tools
- Data-oriented
  - Look at the cases translated by systems
  - Not just automatic metric
- System-oriented
  - Building high-performance, reliable, easy-to-maintain system

Why is MT difficult?


鴨
WHOLEY
Premium Quality
ALL NATURAL
WHOLE
DUCKLING
U.S.D.A. Grade A
兎
WHOLEY
USDA Inspected
Young Tender
WHOLE
Farm Raised
RABBITS
1½ - 2 pound average
Recipes Available
低音
WHOLEY
Pennsylvania
LIVE
Fresh Water
STRIPED
BASS
豬肉
FRESH
PORK
SPARE
RIBS
Great On The Grill!
豬肉
FRESH
PORK
Feet * Tails * Hearts
Fresh Hocks
Liver * Kidney
Neckbones
Stomach * Sides
龍蝦
WHOLEY
Wild Caught!
Live
MAINE
LOBSTERS

Why is MT challenging?

- Ambiguous word boundary
- Polysemy

  He deposited money in a bank account with a high interest rate.

  Sitting on the bank of the Mississippi, a passing ship piqued his interest.

- New entity names
  - COVID-19
- Complex structure
- Ellipsis (i.e. omission)

New Terms

周四经济数据面，美国劳工部报告称，截至8月28日当周美国首次申请失业救济人数为34万，降至2020年美国新冠疫情危机爆发以来的最低点。市场预计该数字为34.5万。

Google Translation (2021.9.1)

On Thursday’s economic data, the U.S. Department of Labor reported that as of August 28, the number of people applying for unemployment benefits for the first time was 340,000, which dropped to the lowest point since the outbreak of the new crown crisis in the United States in 2020. The market expects the number to be 345,000.

VolcTrans (2021.9.1)

On Thursday's economic data, the U.S. Labor Department reported that the number of first-time jobless claims in the United States for the week ending August 28 was 340 thousand, falling to the lowest level since the COVID-19 Epide COVID-19 epidemic crisis broke out in the United States in 2020. The market expects the number to be 345 thousand.

New Terms

周四经济数据面，美国劳工部报告称，截至8月28日当周美国首次申请失业救济人数为34万，降至2020年美国新冠疫情危机爆发以来的最低点。市场预计该数字为34.5万。

Bing Translation (2021.9.1)

On Thursday, the *Labor Department reported that 340,000 people applied for * unemployment benefits for the week ended Aug. 28, the lowest level since the * crisis began in 2020. The market expects the figure to be 345,000.

DeepL (2021.9.1)

On Thursday's economic data front, the U.S. Labor Department reported that the number of first-time U.S. jobless claims for the week ended Aug. 28 was 340,000, falling to the lowest point since the outbreak of the new U.S. crown epidemic crisis in 2020. The market expected the figure to be 345,000.

Complex dependency

周四美股成交额冠军苹果(153.65, 1.14, 0.75%)公司收高0.75%，报153.65美元，创历史收盘新高，成交108.9亿美元，市值逼近2.54万亿美元。

Bing Translation (2021.9.1)

U.S. stock market champion Apple Inc (153.65, 1.14, 0.75 percent) closed up 0.75 percent at $153.65 on Thursday, a record closing high of $10.89 billion, giving it a market capitalization of nearly $2.54 trillion.

DeepL (2021.9.1)

Thursday's U.S. stock turnover leader Apple (153.65, 1.14, 0.75%) closed 0.75% higher at $153.65, an all-time closing high, with $10.89 billion traded and a market cap approaching $2.54 trillion.

他的爷爷和奶奶没见过他的姥姥和姥爷。

Google Translate: His grandpa and grandma have never met his grandma and grandpa.

Correct: His father’s parents never met his mother’s.

Acronym and incorrect word segmentation

一些立陶宛人士表示，中立关系恶化，影响最大的当属立陶宛的出口企业。

Google Translate: Some Lithuanians said that the deterioration of Sino-Lithuanian relations has affected Lithuanian export companies the most.

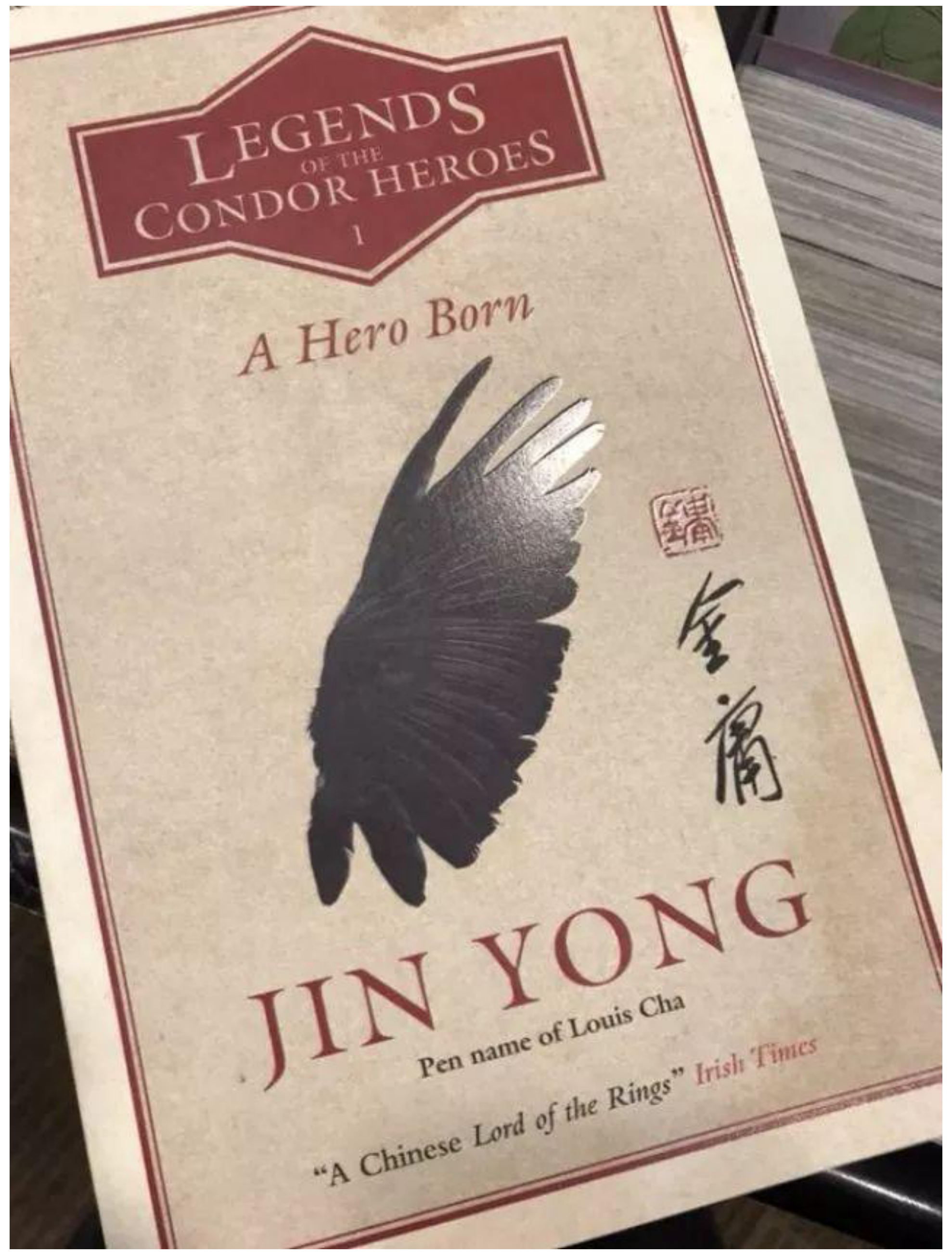
Bing Translate: Some Lithuanians say the deterioration in neutral relations has affected Lithuania's exporters the most.

Made-up Names

- Name:
  - 梅超风 -> Cyclone Mei
  - 王重阳 -> double sun Wang Chongyang
  - Optimus Prime => 擎天柱 or 柯博文
- e.g. made-up martial arts movements

降龙十八掌

the 18 palm attacks to defeat dragons

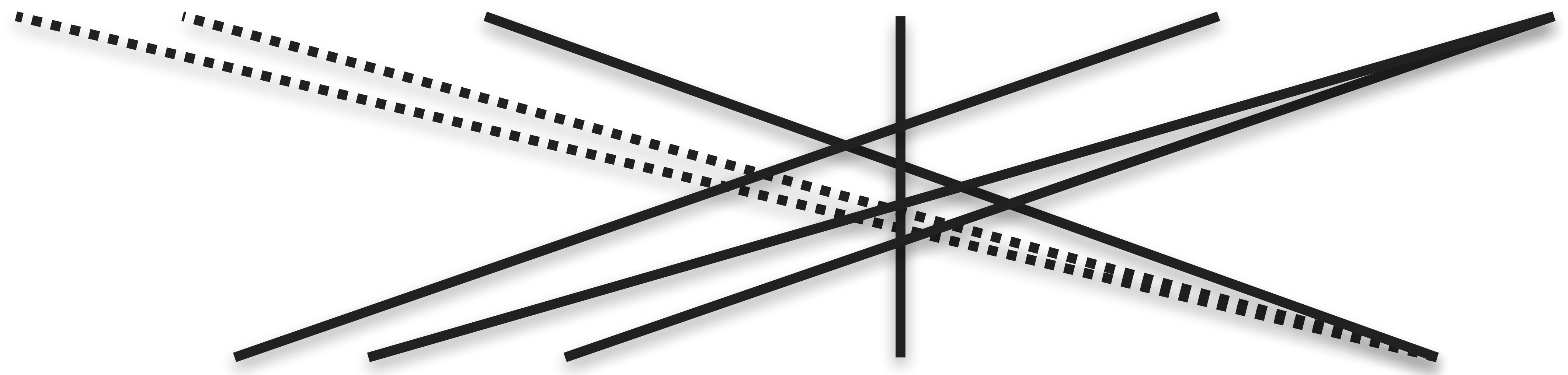


Why is it difficult to translate?

- Structural divergences
  - Morphology
  - Syntax

in the in-city exploded a car-bomb

German: In der Innenstadt explodierte eine Autobombe



English: A car bomb exploded downtown.

Translationese: In the inner city, there exploded a car bomb.

Culture and Slang

这个人很牛

MT1/MT3: This person is very cattle.

MT2: This man is a cow.

MT4: This guy's good.

MT0: This guy is awesome.

Robustness

○ variation of auxiliary function words or symbols

这个人很牛
MT1: This person is very cattle.
MT3: This person is very cattle.
MT0: This guy is awesome.

这个人非常牛。
MT1: This person is very cattle.
MT3: This person is very cattle.
MT0: This guy is awesome.

这个人很牛。
MT1: This person is very bullish.
MT3: This man is very good.
MT4: This guy is good.
MT0: This guy is very good.

这个人很牛!
MT1: This person is very cow!
MT3: This man is very good.
MT4: This man is good!
MT0: This guy is awesome!

Robustness

乔丹最早周日伤愈复出

MT0: Jordan came back from his first injury on Sunday.

MT1: Jordan first recovered from injury on Sunday

乔丹最早周日伤愈复出。

MT0: Jordan came back from injury on Sunday.

MT1: Jordan returned from injury on Sunday.

Reference: Jordan may return from injury as early as this Sunday.

MT: From fluency to nativeness

No, Scarlett, the seeds of greatness were never in me.

MT1: 不，思嘉，伟大的种子永远不会在我身上。

MT0: 不，思嘉，伟大的种子从来就不存在。

Ref: 不，斯佳丽，我根本就不是当大人物的料。

Beauty, Rhythm, Melody in Translation

《黄鹤楼送孟浩然之广陵》

李白

故人西辞黄鹤楼 (lou)，

烟花三月下扬州 (zhou)。

孤帆远影碧空尽，

唯见长江天际流 (liu)。

"The Yellow Crane Tower Sends Meng Haoran's Guangling"

Li Bai

The old man resigned from the Yellow Crane Tower in the west,

Fireworks go down to Yangzhou in March.

A lonely sail is far away and the sky is blue,

Only see the Yangtze River skyline flow.

(by VolcTrans)

Beauty, Rhythm, Melody in Translation

《黄鹤楼送孟浩然之广陵》

李白

故人西辞黄鹤楼 (lou),

烟花三月下扬州 (zhou)。

孤帆远影碧空尽,

唯见长江天际流 (liu)。

"Celestial Crane Pavilion: Sending Meng Haoran Off to Guangling”

Li Bai

Old friends westward part from Yellow Crane Tower,

Amidst March's mist, they descend to Yangzhou's bower.

Lone sail, distant shadow, in the azure vast,

Only the Yangtze River stretches far at last.

(by ChatGPT)

Beauty, Rhythm, Melody in Translation

《黄鹤楼送孟浩然之广陵》

李白

故人西辞黄鹤楼 (lou)，

烟花三月下扬州 (zhou)。

孤帆远影碧空尽，

唯见长江天际流 (liu)。

Seeing Meng Hao-ran Off At Yellow Crane Tower

by Li Bai

My friend has left the west where the Yellow Crane towers;

For River Town green with willows and red with flowers.

His lessening sail is lost in the boundless blue sky;

Where I see but the endless River rolling by.

(translated by Xu Yuanchong)

(Average) Human Level Translation

You say that you love rain, but you op
en your umbrella when it rains.
You say that you love the sun, but you
find a shadow spot when the sun shi
nes.
You say that you love the wind, but yo
u close your windows when wind blo
ws.
This is why I am afraid, you say that y
ou love me too.

MT: 你说你喜欢雨，但雨下的时候你打开雨伞。

你说你爱太阳，但当太阳照耀时，你发现了一个阴影斑点。

你说你喜欢风，但是当风吹起的时候你会关上窗户。

这就是为什么我害怕，你说你也爱我。

Expert Level Translation

诗经体：

子言慕雨，启伞避之。子言好阳，寻荫拒之。子言喜风，阖户离之。子言偕老，吾所畏之。

离骚版:

君乐雨兮启伞枝，君乐昼兮林蔽日，君乐风兮栏帐起，君乐吾兮吾心噬。

七律：

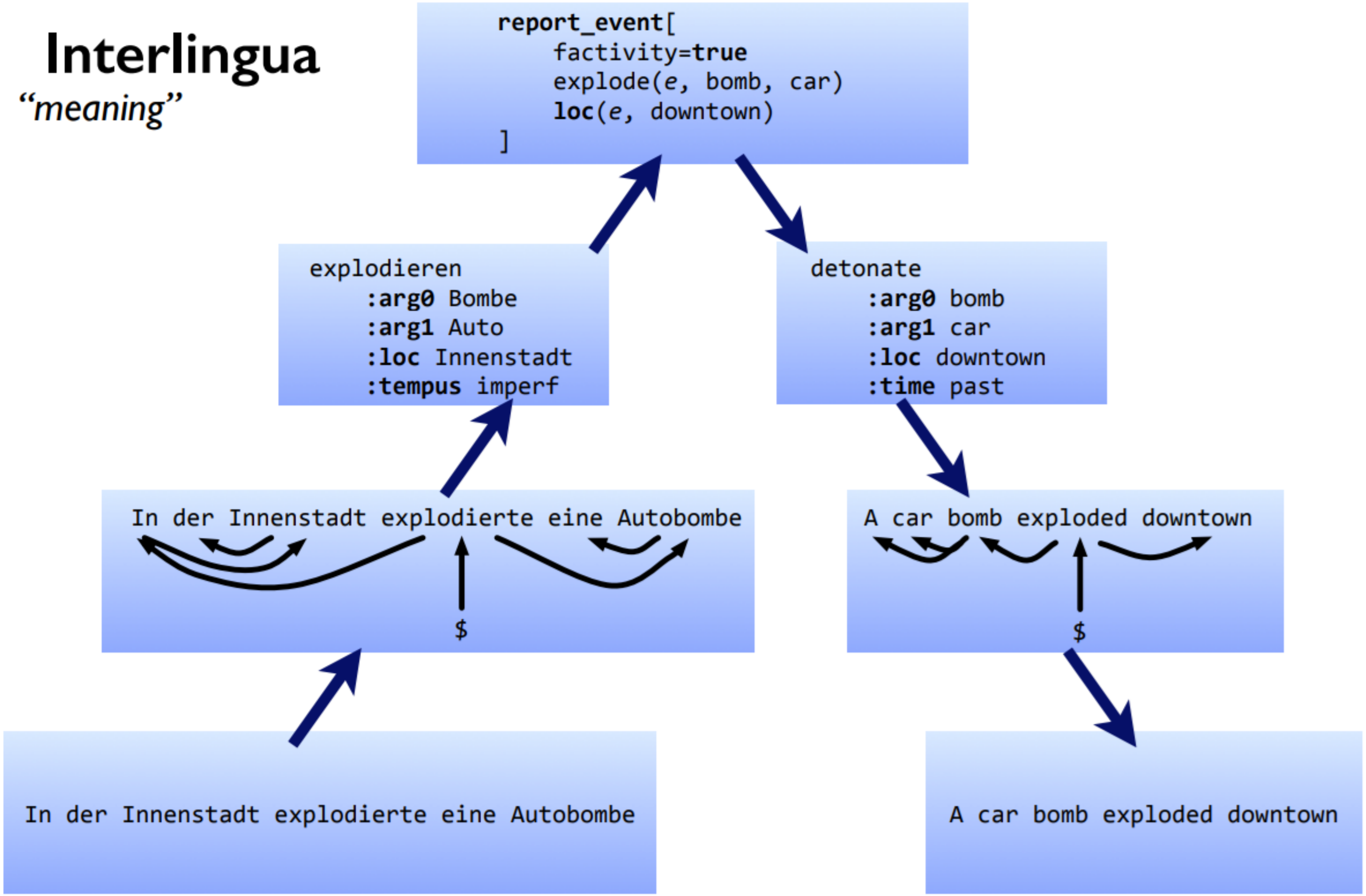
江南三月雨微茫，罗伞叠烟湿幽香。夏日微醺正可人，却傍佳木趁荫凉。霜风清和更初霁，轻蹙蛾眉锁朱窗。怜卿一片相思意，犹恐流年拆鸳鸯。

网络咆哮体：

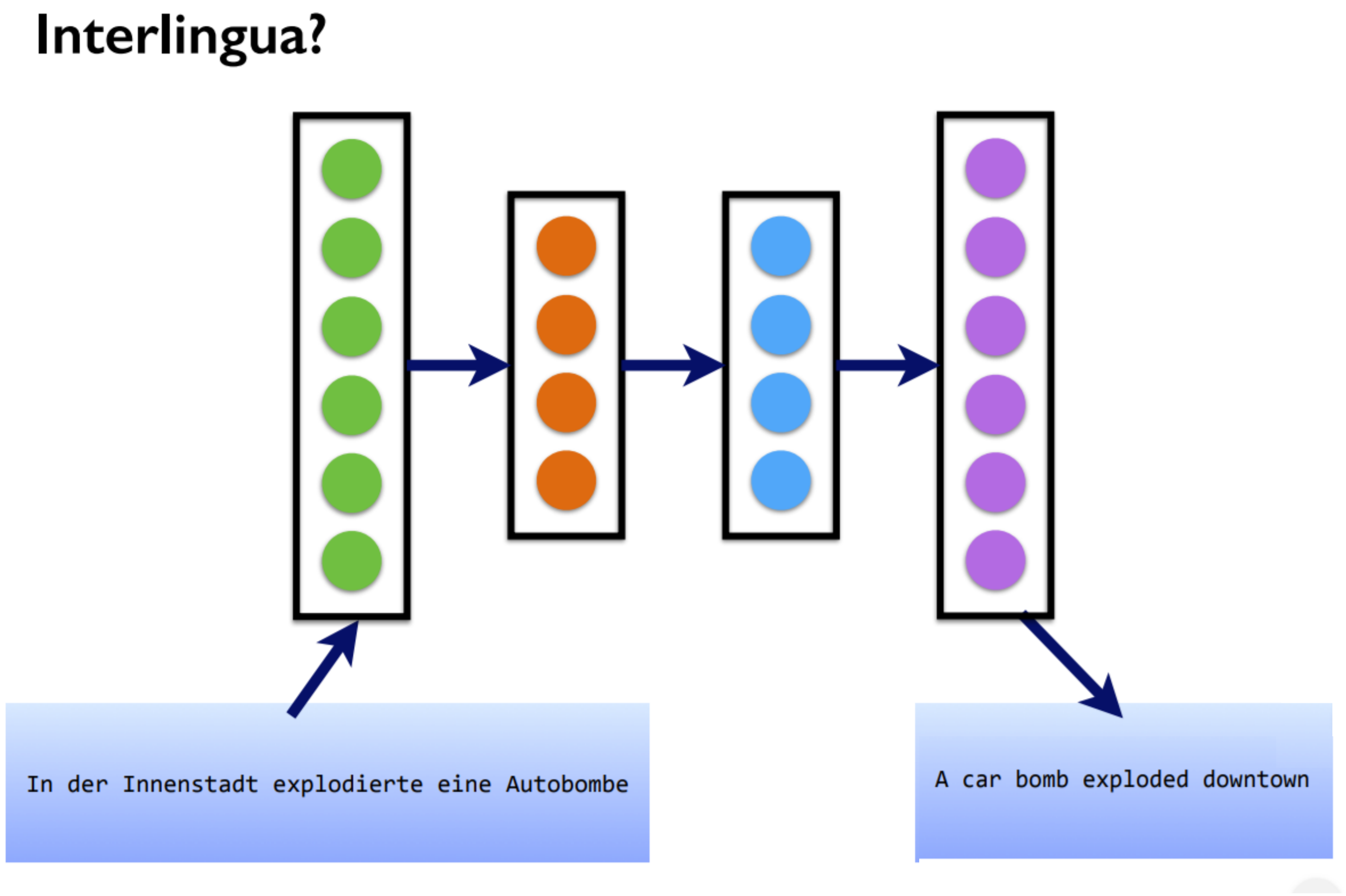
你有本事爱雨天，你有本事别打伞啊！你有本事爱阳光，你有本事别乘凉啊！！你有本事爱吹风，你有本事别关窗啊！！！你有本事说爱我，你有本事捡肥皂啊！！！！

Mathematical Framework of MT

Finding Interlingua for Translation



Interlingua can be implicit representation



Mathematical Frameworks of MT

direct conditional probabilistic translation model

$$\operatorname{argmax} p_\theta(y \mid x)$$

Autoregressive Translation

$$p_\theta(y \mid x) = \prod_i p(y_i \mid x, y_{1:i-1})$$

Non-Autoregressive Translation

$$p_\theta(y \mid x) = \prod_i p(y_i \mid x)$$

$$\operatorname{argmax} p_\theta(y \mid x) \propto p(x \mid y) p(y)$$

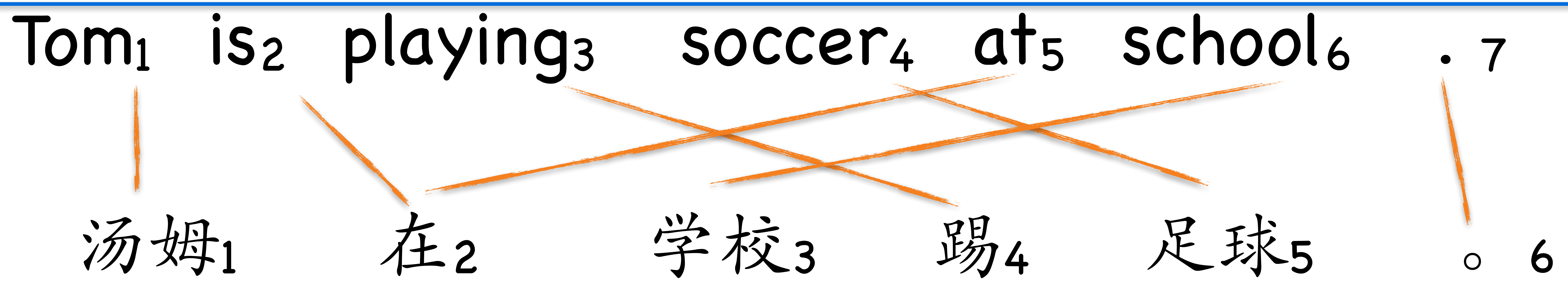
IBM model

reverse translation probability   language model

Tom is playing soccer at school .

汤姆 $\left\{ \begin{array}{l} \text{在0.4} \\ \text{是0.3} \end{array} \right.$ 学校 $\left\{ \begin{array}{l} \text{踢0.3} \\ \text{玩0.4} \end{array} \right.$ 足球 。

Statistical MT



- The translation prob. $p(x \mid y) = \sum_{a} f(x, a \mid y)$
- IBM model 1:

$$p(x \mid y) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_l=0}^{m} \prod_{j=1}^{m} t(x_j \mid y_{a_j})$$

Statistical MT

- The first model IBM Model 1 is over simplified with a lot of independence assumptions

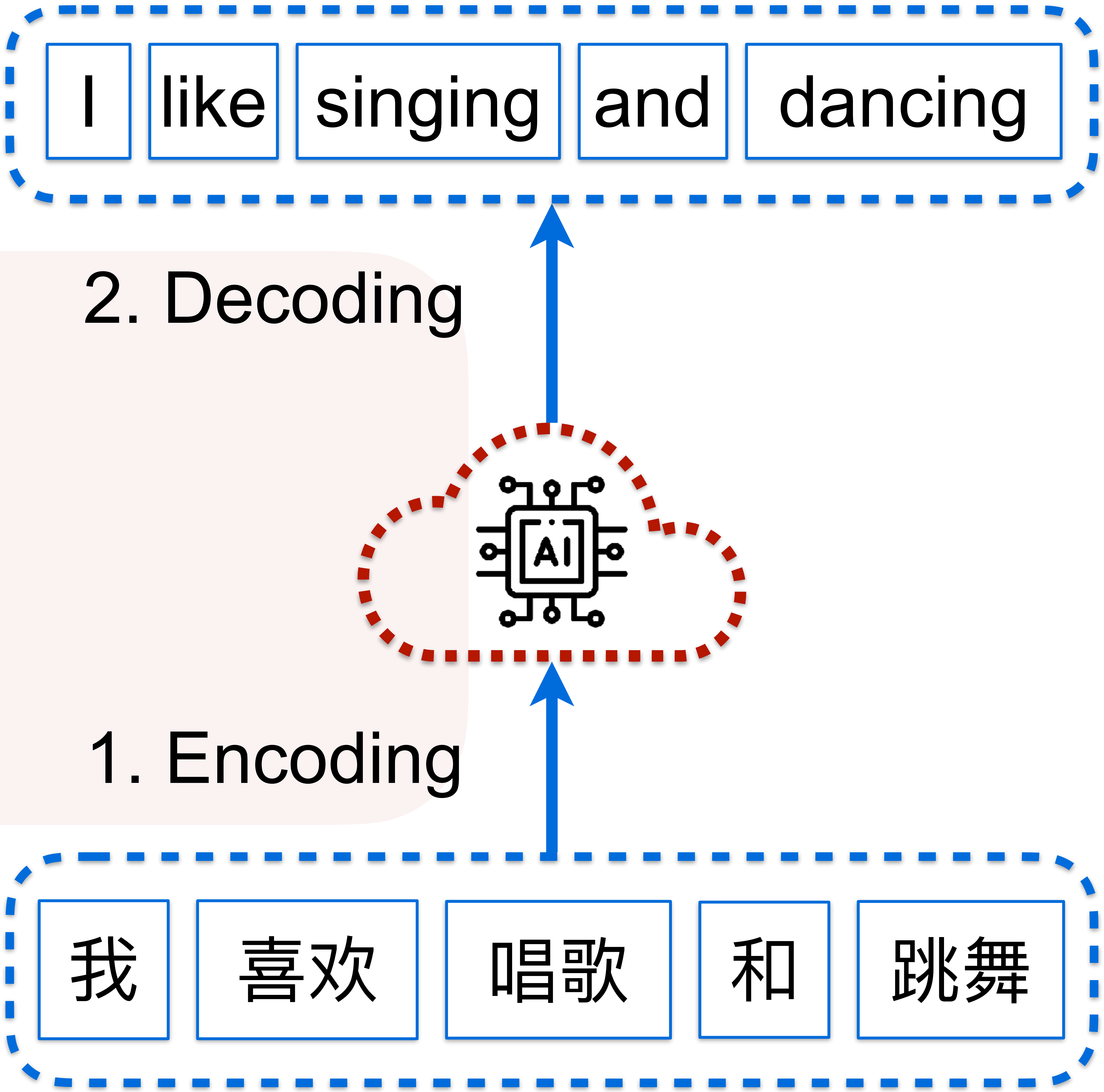
IBM Model 1	Lexical model
IBM Model 2	global alignment model, alignment is dependent on position
IBM Model 3	adding fertility model
IBM Model 4	relative reordering model
IBM Model 5	fixes deficiency

Peter Brown et al, The mathematics of statistical machine translation: Parameter estimation, 1993.

Neural Machine Translation

Transformer Model

$$p_\theta(y \mid x) = \prod_i p(y_i \mid x, y_{1:i-1})$$



MT Data

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)

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

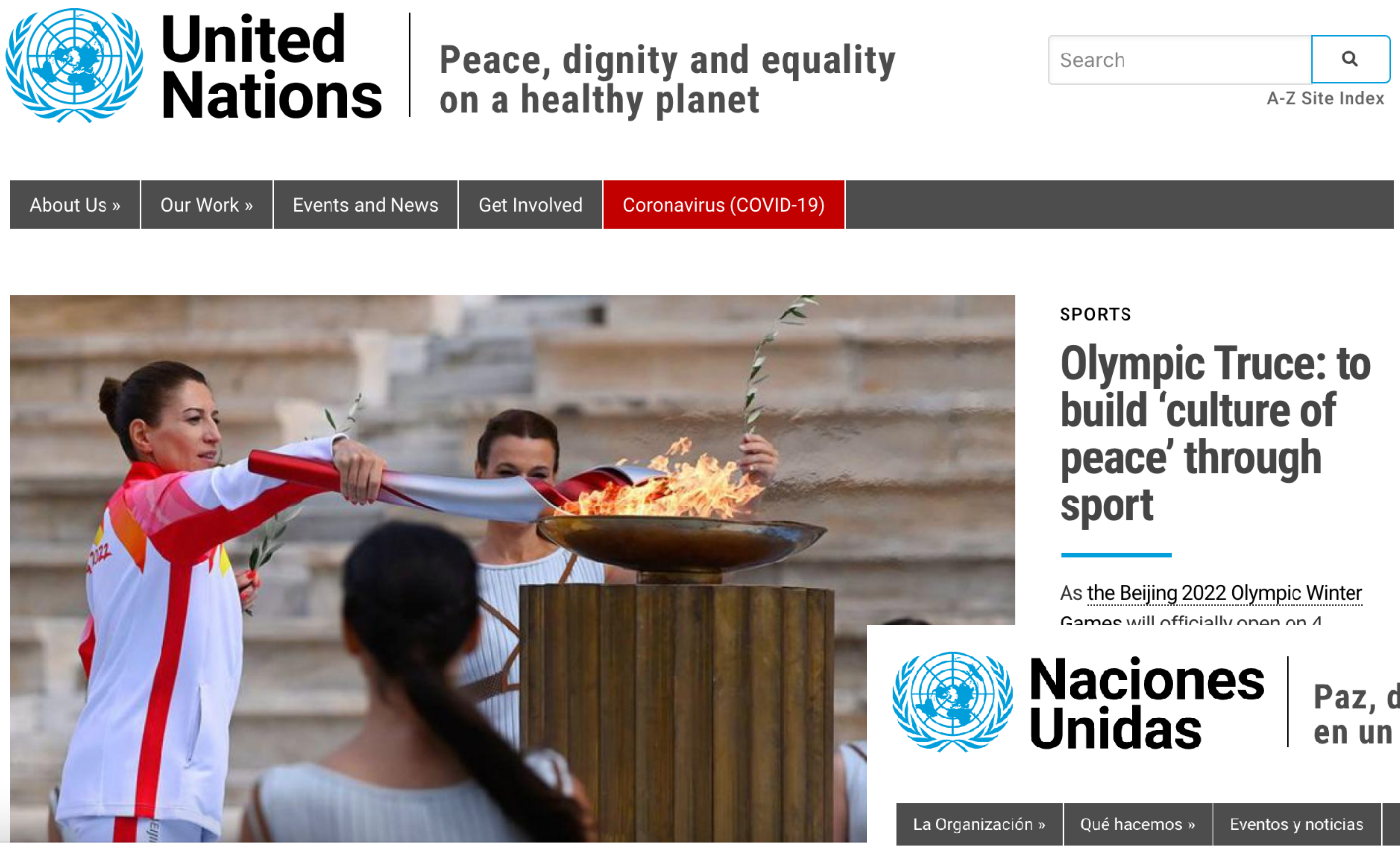
Some Projects using OPUS

- Let'sMT! - On-line SMT toolkit

Sub-corpora (downloads & infos):

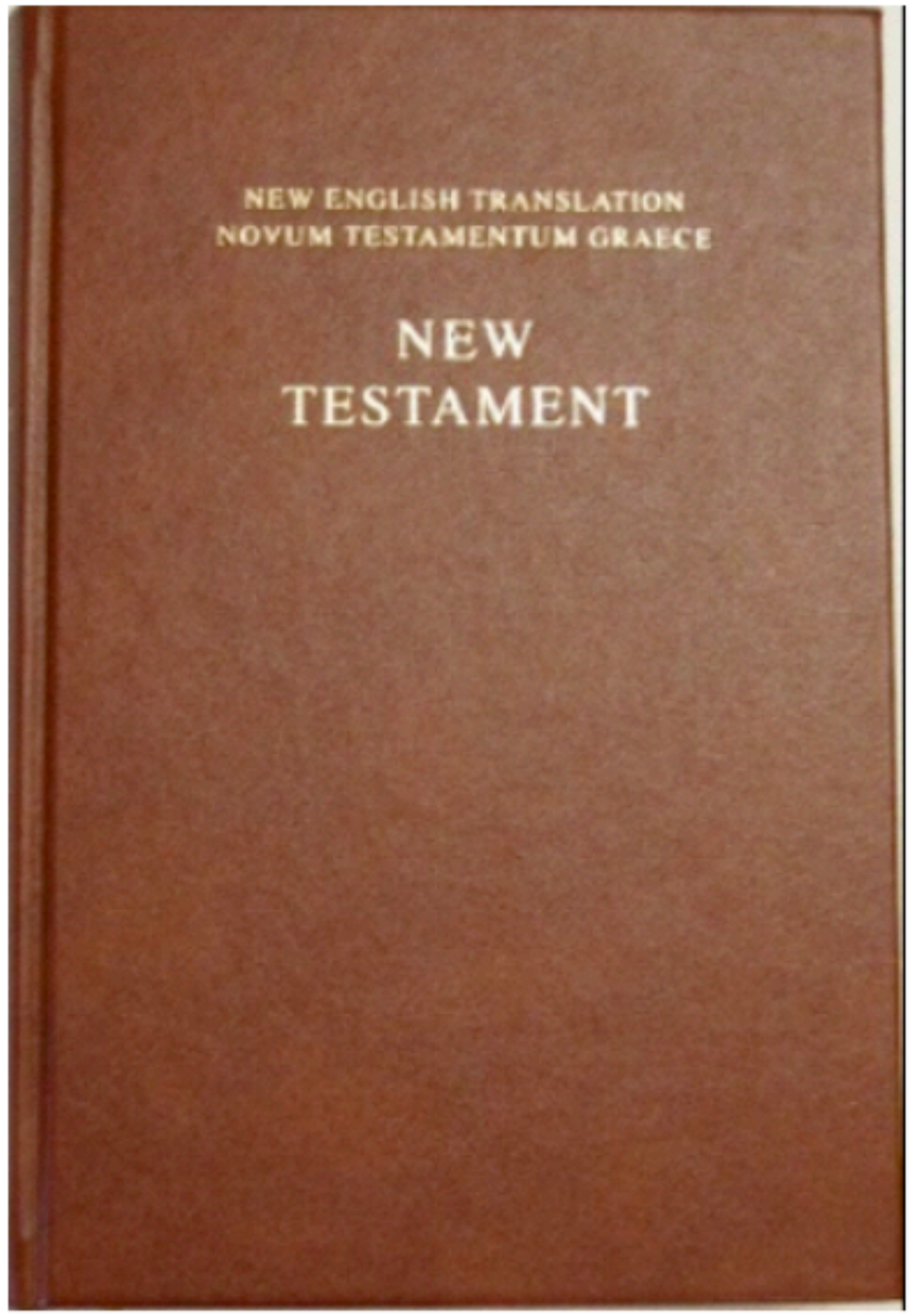
- Books - A collection of translated literature (Books.tar.gz - 535 MB)
- DGT - A collection of EU Translation Memories provided by the JRC
- DOGC - Documents from the Catalan Goverment (DOGC.tar.gz - 2.8 GB)
- ECB - European Central Bank corpus (ECB.tar.gz - 3.0 GB)
- EMEA - European Medicines Agency 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)

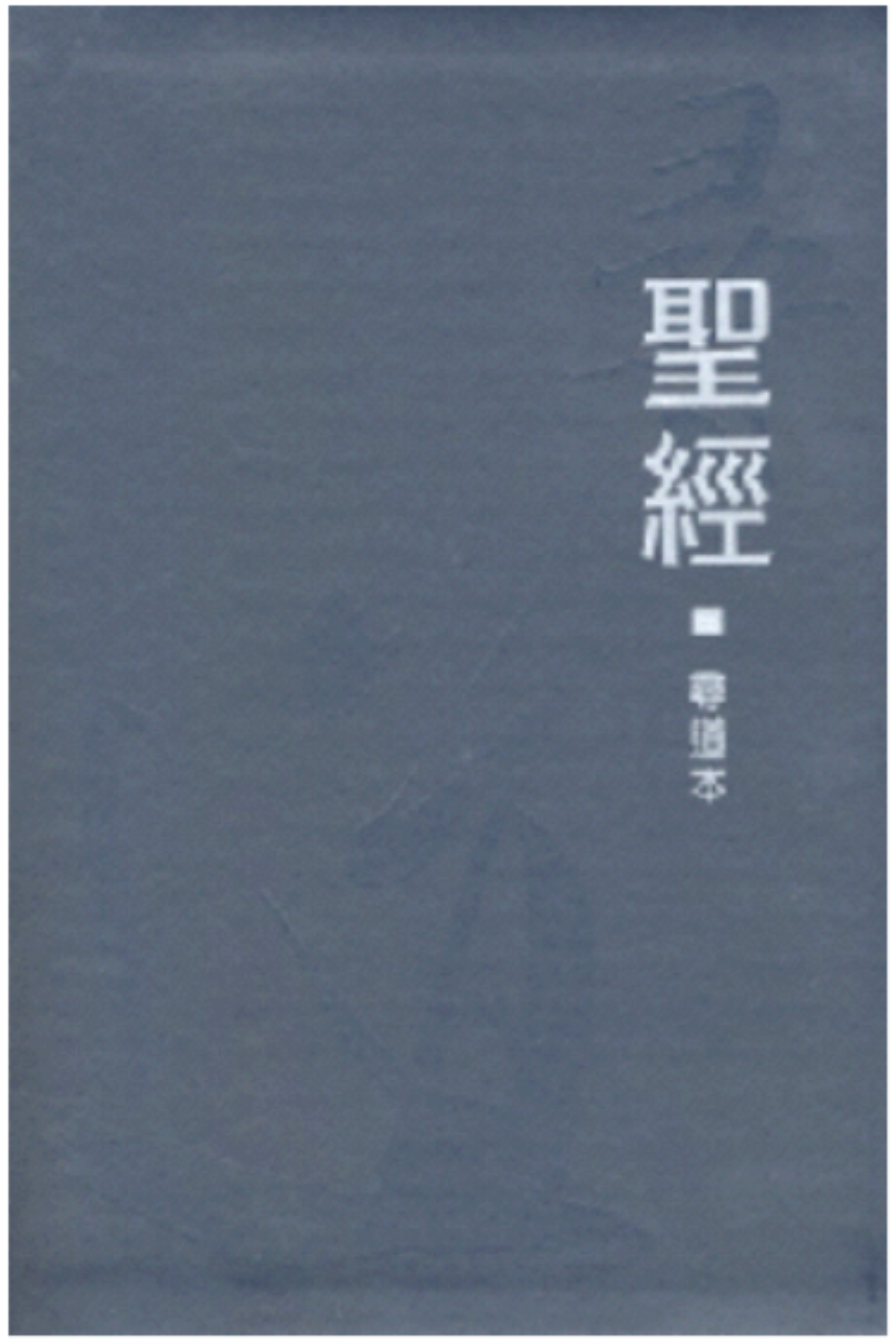
Parallel corpora



Parallel corpora

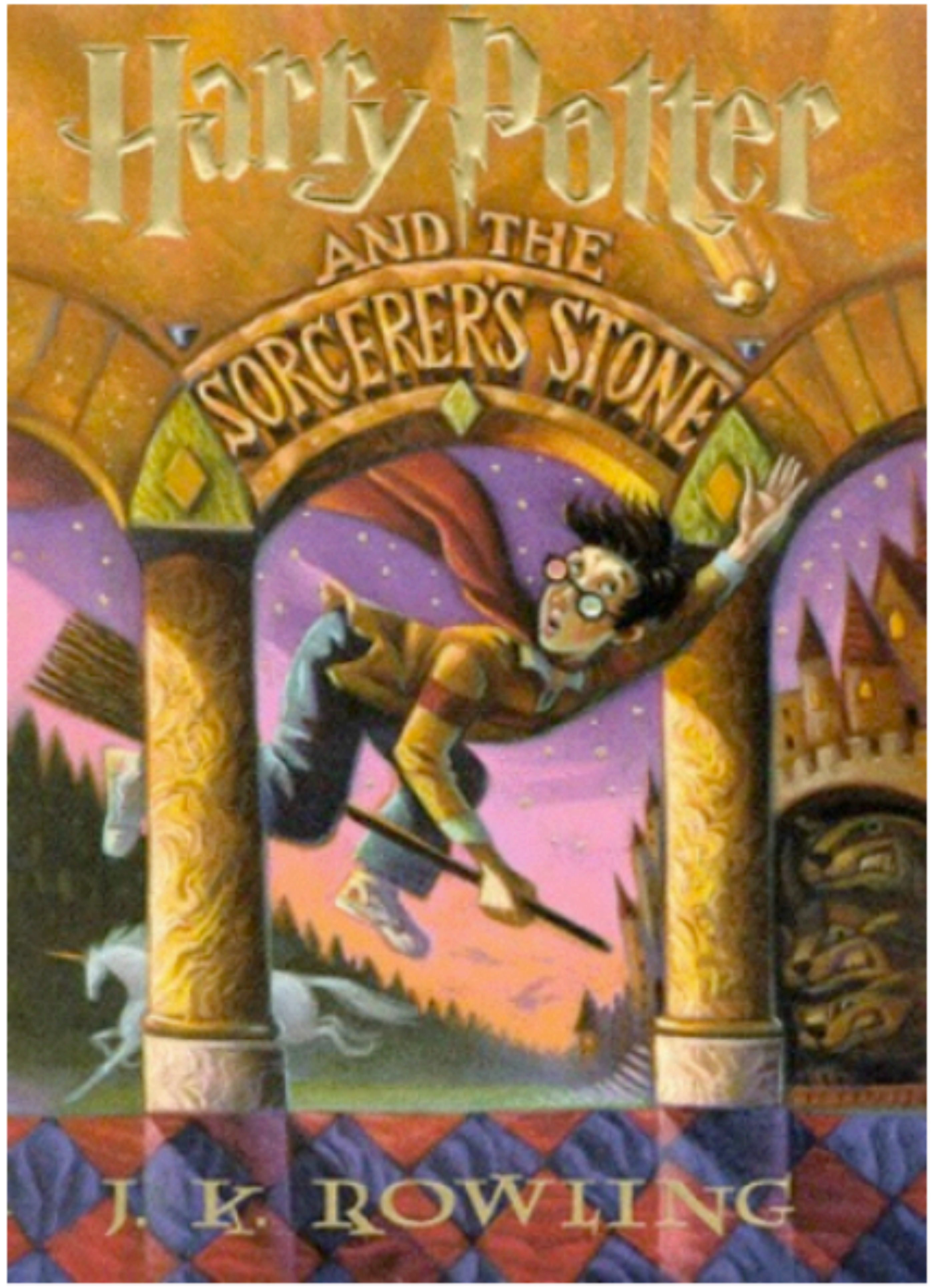
- Translation in 724 languages
- A portion: in 3,589 languages

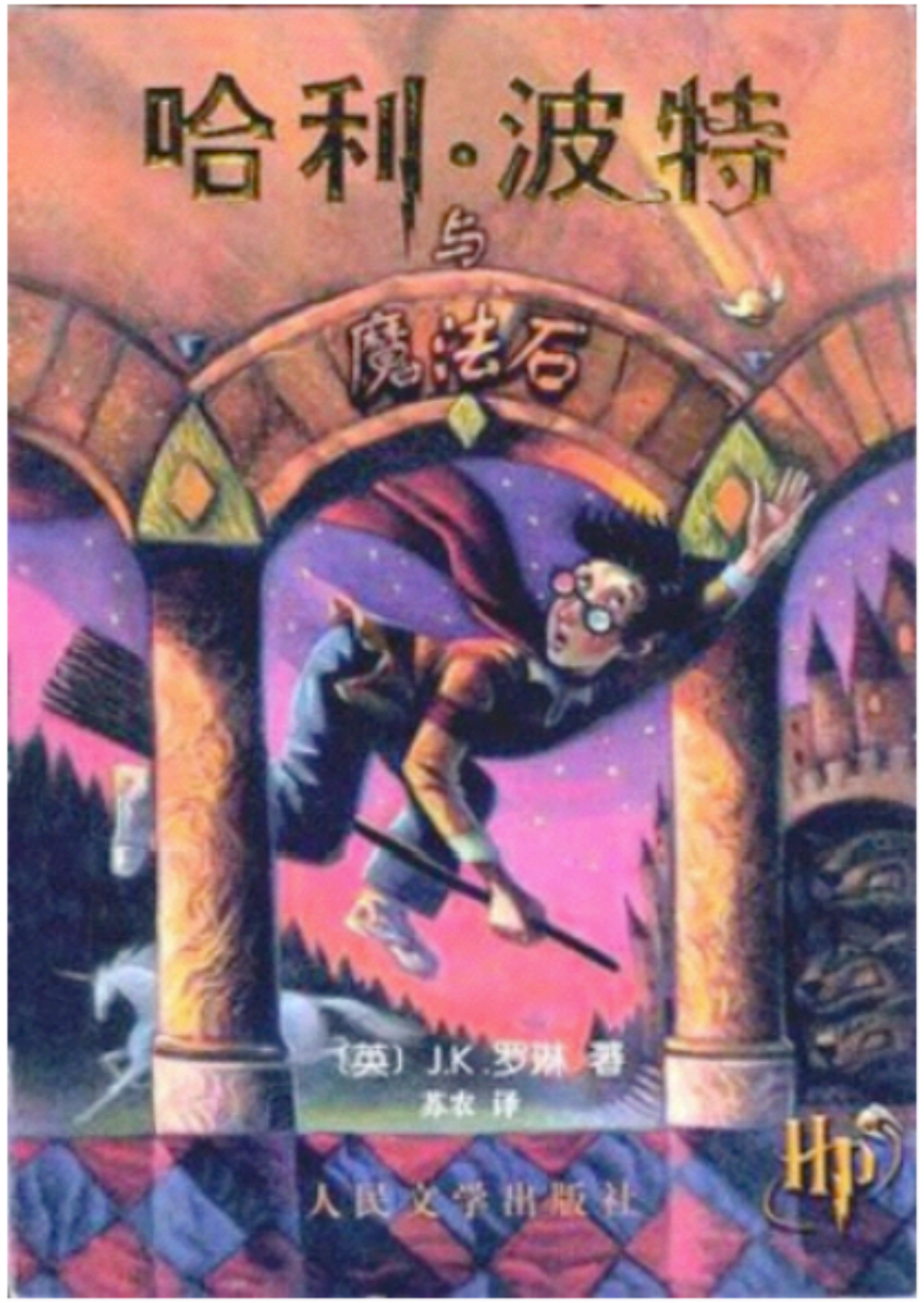




Parallel corpora

Popular books in multiple languages





Data Resources

- (Text) Machine Translation:
  - News Translation (general domain): http://statmt.org/wmt23/translation-task.html
    - includes data from many sources:
      - Europarl
      - UN Parallel Corpus
  - OPUS: https://opus.nlpl.eu/index.php
- Speech Translation:
  - MuST-C: https://ict.fbk.eu/must-c/
  - CoVoST: https://github.com/facebookresearch/covost
  - LibriSpeech
- Tatoeba: collections of translations, https://tatoeba.org/en
- Wikipedia: raw corpus
- Common-crawl: a very large dataset of crawled web pages, noisy.

Mining Parallel Corpus from the Web

- Usually start from a subset of common-crawl.
- Using bilingual sentence embeddings to filter possible parallel sentences
- e.g. Laser embedding (90 languages)
- Usually only filter candidates within a same page.
- Could be costly
- ccmatrix dataset (created by Meta)

MT Evaluation

Many possible translation, which is better?

SpaceX周三晚间进行了一次发射任务，将四名毫无航天经验的业余人士送入太空轨道。

SpaceX launched a mission Wednesday night to put four amateurs with no space experience into orbit.

SpaceX conducted a launch mission on Wednesday night, sending four amateurs with no aerospace experience into space orbit.

SpaceX conducted a launch mission Wednesday night that sent four amateurs with no spaceflight experience into orbit.

SpaceX carried out a launch mission on Wednesday night to put four amateurs without Aerospace experience into orbit.

Assessing the Quality of Translation

- Criteria for evaluation metric
  - Consistent across different evaluation, so that translation quality is comparable
  - Differentiable: tell high quality translation from low quality ones
  - Low cost: requires low effort of human (e.g. amateur can perform) or computation

Aspects of Translation Quality

- Intuition
  - Scoring of translations is (implicitly) based on an identification of errors and other imperfections.
- Adequacy/Faithfulness
  - Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?
- Expressiveness
- Elegance
- Due to Yan Fu (1854-1921)

Direct Assessment of Translation Quality

- Source-based
    - Human annotators are given source, without reference.
    - avoid bias
    - can also be used to evaluate human translation performance
- Reference-based
    - Human annotators are given reference, without source.
    - Can be done by monolingual speaker in target language
    - Less effort
- Source-Reference

Direct Assessment of Translation Quality

- Grading scheme
  - 1-4, 1-5, 1-6
  - 0-100 scale (used in WMT 2020)
- Does it require professional translator or amateur(college students in Foreign language)

4	Correct translation and fluent language
3	Mostly understandable, with 1 or 2 errors
2	some meaningful, but more errors
1	incorrect or major errors

WMT 2020 Evaluation

- 2887 Turkers recruited on Amazon Mechanical Turk.
- 2233 are removed, not passing the quality control
- 654 Turkers are adopted
- 166,868 assessment scores (of 654k)
- For 10 to-English pairs (Chinese, Czech, German, Russian, etc.)
- Turkers are provided source and machine translated output
- Quality Control (next)

Barrault et al. Findings of the 2020 Conference on Machine Translation (WMT20), 2020

Quality Control

- How to ensure that crowd raters produce high quality assessment?
- 100 translation assessment: 40 are regular
- Repeat pairs (10): expecting similar judgement
- Bad Reference Pairs (10):
  - damaged MT outputs by randomly replacing n-gram phrases from the same test set.
  - expects low scores
- Good Reference Pairs (10)
  - Use golden reference
  - expects high scores
- Excluding Bad (10) and Good (10) in calculating final score.

Filtering Low-quality Annotators

- How to tell if an annotator consistently scores bad references pairs lower?
- Hypothesis testing (significance test)
  - Annotator scores MT pair with X
  - Annotator scores Bad Reference Pair Y
  - Y < X
  - Is the annotator reliable in assessment? (Is the difference statistically significant?)
- Remove annotators whose scores for normal MT not different from bad reference pairs!

Is the score of system A better than B?

- n pairs of (e.g. MT output, degraded bad translation)
- Scores from human annotators for each (x_i, y_i)
- Null Hypothesis:

ui=xi - yi is close to 0

- Test statistic:

$t = \dfrac{\bar{u}}{s/\sqrt{n}}$, where mean difference $\bar{u} = \dfrac{u_i}{n} = \dfrac{x_i - y_i}{n}$,

standard deviation: $s = \sqrt{\dfrac{1}{n-1}(u_i - \bar{u})^2}$

- e.g. WMT20, n is 10 (for one 100-item batch)
- Compare with t-distribution table: T=1.645 for p-value 0.05

Alternative Annotator Agreement

- For discrete scores (e.g. 1-4)
- Kappa coefficient

$$\kappa = \frac{p(A) - p_r}{1 - p_r}$$

- p(A): percentage of agreed assessments
- p_r: percentage of agreement if random guess (=1/K if there K discrete labels)
- e.g. P(A) = 0.4, P_r=0.25, k=0.2

Ranking and Annotator Difference

- In WMT20, scores of a same annotators are normalized by according to mean and standard deviation
- The overall score is an average of standardized scores.
- Ranking based on overall-score (avg z)

Example Results from WMT 20

Chinese→English		
Ave.	Ave. z	System
77.5	0.102	VolcTrans
77.6	0.089	DiDi-NLP
77.4	0.077	WeChat-AI
76.7	0.063	Tencent-Translation
77.8	0.060	Online-B
78.0	0.051	DeepMind
77.5	0.051	OPPO
76.5	0.028	THUNLP
76.0	0.016	SJTU-NICT
72.4	0.000	Huawei-TSC
76.1	−0.017	Online-A
74.8	−0.029	HUMAN
71.7	−0.071	Online-G
74.7	−0.078	dong-nmt
72.2	−0.106	zlabs-nlp
72.6	−0.135	Online-Z
67.3	−0.333	WMTBiomedBaseline

English→Chinese		
Ave.	Ave. z	System
80.6	0.568	HUMAN-B
82.5	0.529	HUMAN-A
80.0	0.447	OPPO
79.0	0.420	Tencent-Translation
77.3	0.415	Huawei-TSC
77.4	0.404	NiuTrans
77.7	0.387	SJTU-NICT
76.6	0.373	VolcTrans
73.7	0.282	Online-B
73.0	0.241	Online-A
69.5	0.136	dong-nmt
68.5	0.135	Online-Z
70.1	0.122	Online-G
68.7	0.082	zlabs-nlp

Example Results from WMT 20

Japanese→English

Ave.	Ave. z	System
75.1	0.184	Tohoku-AIP-NTT
76.4	0.147	NiuTrans
74.1	0.088	OPPO
75.2	0.084	NICT-Kyoto
73.3	0.068	Online-B
70.9	0.026	Online-A
71.1	0.019	eTranslation
64.1	−0.208	zlabs-nlp
66.0	−0.220	Online-G
61.7	−0.240	Online-Z

English→Japanese

Ave.	Ave. z	System
79.7	0.576	HUMAN
77.7	0.502	NiuTrans
76.1	0.496	Tohoku-AIP-NTT
75.8	0.496	OPPO
75.9	0.492	ENMT
71.8	0.375	NICT-Kyoto
71.3	0.349	Online-A
70.2	0.335	Online-B
63.9	0.159	zlabs-nlp
59.8	0.032	Online-Z
53.9	−0.132	SJTU-NICT
52.8	−0.164	Online-G

Example Results from WMT 20

German→English

Ave.	Ave. z	System
82.6	0.228	VolcTrans
84.6	0.220	OPPO
82.2	0.186	HUMAN
81.5	0.179	Tohoku-AIP-NTT
81.3	0.179	Online-A
81.5	0.172	Online-G
79.8	0.171	PROMT-NMT
82.1	0.167	Online-B
78.5	0.131	UEDIN
78.8	0.085	Online-Z
74.2	−0.079	WMTBiomedBaseline
71.1	−0.106	zlabs-nlp
20.5	−1.618	yolo

English→German

Ave.	Ave. z	System
90.5	0.569	HUMAN-B
87.4	0.495	OPPO
88.6	0.468	Tohoku-AIP-NTT
85.7	0.446	HUMAN-A
84.5	0.416	Online-B
84.3	0.385	Tencent-Translation
84.6	0.326	VolcTrans
85.3	0.322	Online-A
82.5	0.312	eTranslation
84.2	0.299	HUMAN-paraphrase
82.2	0.260	AFRL
81.0	0.251	UEDIN
79.3	0.247	PROMT-NMT
77.7	0.126	Online-Z
73.9	−0.120	Online-G
68.1	−0.278	zlabs-nlp
65.5	−0.338	WMTBiomedBaseline

Example Results from WMT 20

German→ French

Ave.	Ave. z	System
90.4	0.279	OPPO
90.2	0.266	VolcTrans
89.7	0.262	IIE
89.2	0.243	HUMAN
89.1	0.226	Online-B
89.1	0.223	Online-A
88.5	0.208	Online-G

French → German

Ave.	Ave. z	System
89.8	0.334	VolcTrans
89.7	0.333	OPPO
89.1	0.319	IIE
89.0	0.295	Online-B
87.4	0.247	HUMAN
87.3	0.240	Online-A
87.1	0.221	SJTU-NICT
86.8	0.195	Online-G
85.6	0.155	Online-Z

Expert Rating - MQM

- Multidimensional Quality Metrics
- Rate with error category and severity level
- Error Category: Accuracy, Fluency, Terminology, Style, and Locale
- -25 to 0

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

Freitag et al, Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation, 2021

MQM Error Category

Error Category		Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize the 5 most severe errors.

Automatic Metric

- The need of automatic metric:
  - Human evaluation is expensive
  - Need fast turnaround for model development
- Easy for text classification, just comparing one label
- Hard for variable-length sequence
  - multiple yet correct translation
- Widely adopted metric: BLEU
  - BiLingual Evaluation Understudy

BLEU

- Measuring the precision of n-grams
  - Precision of n-gram: percentage of tokens in output sentences

$$\_\, p_n = \frac{num.of.correct.token.ngram}{total.output.ngram}$$

- Penalize for brevity
  - if output is too short
  - $bp = min(1, e^{1-r/c})$
- BLEU=$bp \cdot (\prod p_i)^{\frac{1}{4}}$
- Notice BLEU is computed over the whole corpus, not on one sentence

Example

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System A: SpaceX launched a mission Wednesday evening into a space orbit.

System B: A rocket sent SpaceX into orbit Wednesday.

Example

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System A: SpaceX launched a mission Wednesday evening into a space orbit.

	Precision
Unigram	9/11
Bigram	4/10
Trigram	2/9
Four-gram	1/8

$$bp = e^{1-12/11} = 0.91$$

$$BLEU = 0.91 * (9/11 * 4/10 * 2/9 * 1/8)^{1/4}$$

$$= 28.1\%$$

Exercise: Calculate BLEU

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System B: A rocket sent SpaceX into orbit Wednesday.

Multi-BLEU

- To account for variability if one source has multiple references.
- Precision
  - n-grams can match in any of the references
  - $p_n = \dfrac{num\,.\,of\,.\,correct\,.\,token\,.\,ngram}{total\,.\,output\,.\,ngram}$
- Brevity Penalty
  - $bp = min(1, e^{1-r/c})$
  - closest reference length used
- BLEU=$bp \cdot (\prod p_i)^{\frac{1}{4}}$
- Notice BLEU is computed over the whole corpus, not on one sentence

Pitfall in Calculating BLEU

- Be careful! Tokenization and normalization make diff!

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System A: SpaceX launched a mission Wednesday evening into a space orbit.

- What is the BLEU for Char-level Tokenization:

  Ref: A S p a c e X r o c k e t w a s l a u n c h e d i n t o a s p a c e o r b i t W e d n e s d a y e v e n i n g .

  System A: S p a c e X l a u n c h e d a m i s s i o n W e d n e s d a y e v e n i n g i n t o a s p a c e o r b i t .

BLEU scores can differ much!

Data from WMT17 for the same system output using different BLEU configuration.

config	English→ ⋆ en-cs	en-de	en-fi	en-lv	en-ru	en-tr	⋆ →English cs-en	de-en	fi-en	lv-en	ru-en	tr-en
basic	20.7	25.8	22.2	16.9	33.3	18.5	26.8	31.2	26.6	21.1	36.4	24.4
split	20.7	26.1	22.6	17.0	33.3	18.7	26.9	31.7	26.9	21.3	36.7	24.7
unk	20.9	26.5	25.4	18.7	33.8	20.6	26.9	31.4	27.6	22.7	37.5	25.2
metric	20.1	26.6	22.0	17.9	32.0	19.9	27.4	33.0	27.6	22.0	36.9	25.6
range	0.6	0.8	0.6	1.0	1.3	1.4	0.6	1.8	1.0	0.9	0.5	1.2
basic_{lc}	21.2	26.3	22.5	17.4	33.3	18.9	27.7	32.5	27.5	22.0	37.3	25.2
split_{lc}	21.3	26.6	22.9	17.5	33.4	19.1	27.8	32.9	27.8	22.2	37.5	25.4
unk_{lc}	21.4	27.0	25.6	19.1	33.8	21.0	27.8	32.6	28.3	23.6	38.3	25.9
metric_{lc}	20.6	27.2	22.4	18.5	32.8	20.4	28.4	34.2	28.5	23.0	37.8	26.4
$\textit{range}_{lc}$	0.6	0.9	0.5	1.1	0.6	1.5	0.7	1.7	1.0	1.0	0.5	1.2

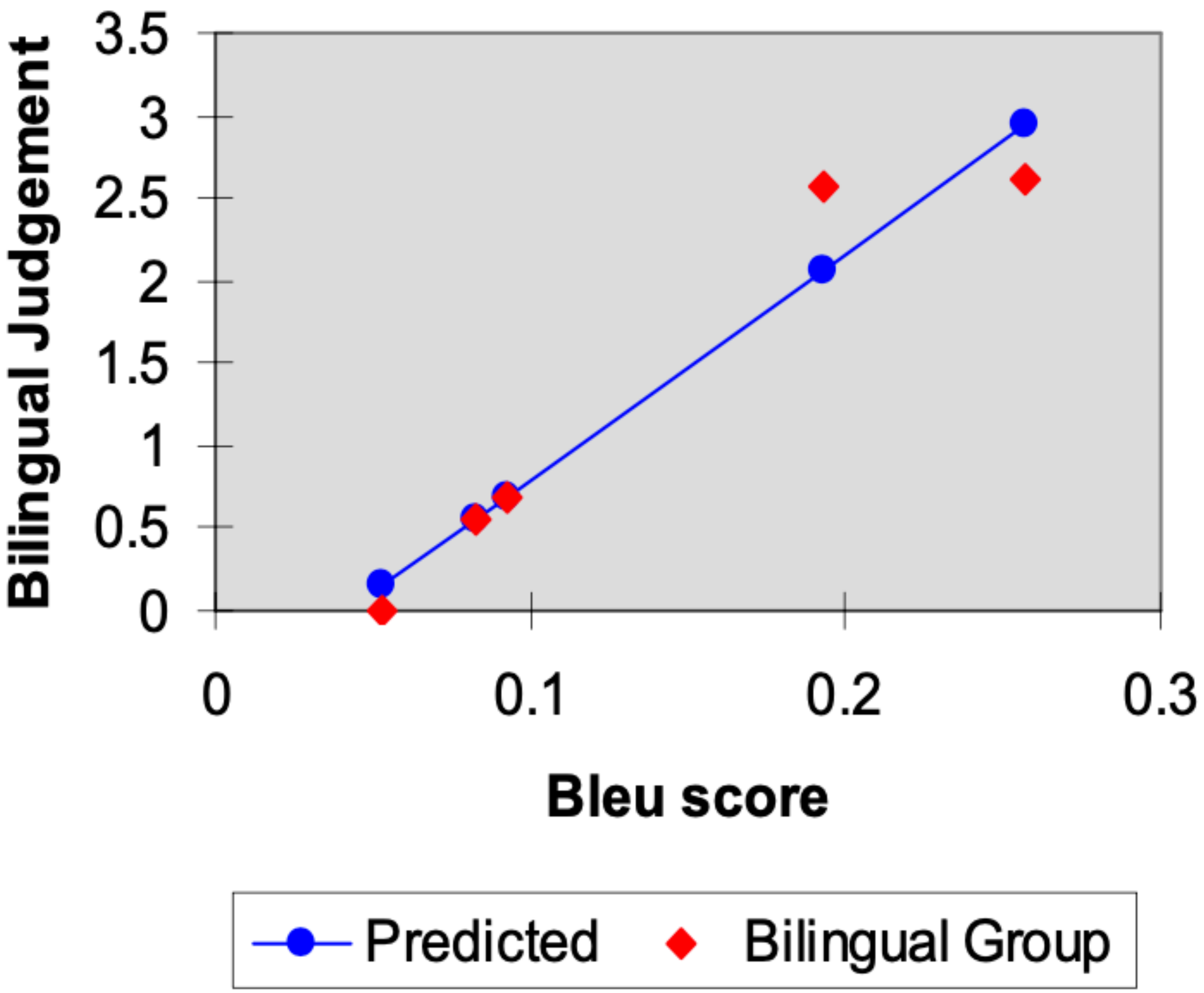
Matt Post. A Call for Clarity in Reporting BLEU Scores, 2018

Guideline of Using BLEU

- Always use sacreBLEU to report
  - also known as detokenized BLEU
  - use metric's original tokenization, no processing on the reference data!!!
    - because different way to tokenize, whether to split compound words (e.g. long-term ==> long - term), cased or uncased can all affect BLEU
- more than 100 languages
  - spBLEU (BLEU with sentence-piece tokenization)
  - warning: can be inflated.

Is BLEU correlated with Human Evaluation?

Figure 6: BLEU predicts Bilingual Judgments



Papenani et al, BLEU: a Method for Automatic Evaluation of Machine Translation. 2002

Learned Metrics

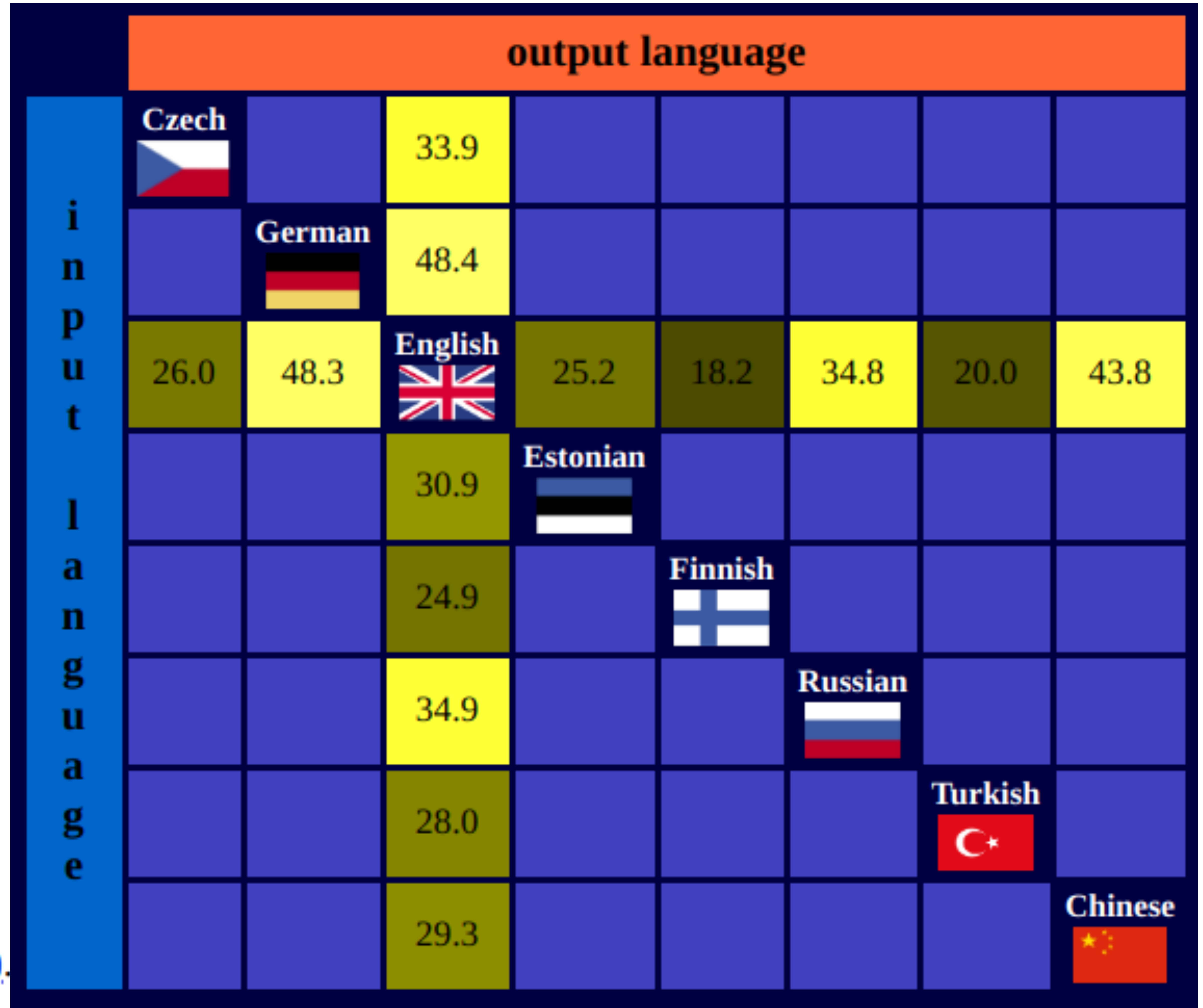
- **Supervised:**
  - **BLEURT:** Train BERT to predict human evaluation scores (Sellam et al. 2020)
  - **COMET:** Train model to predict human eval, also using source sentence (Rei et al. 2020)
- **Unsupervised/Semi-supervised**
  - **SEScore & SEScore2:** synthesize MQM style errors and train (Xu et al 22&23)
  - **BertScore:** Find similarity between BERT embeddings (unsupervised) (Zhang et al. 2020)
  - **PRISM:** Model based on training paraphrasing model (Thompson and Post 2020)
  - **BARTScore:** Calculate the probability of source, reference, or system output (Yuan et al. 2021)

Which One to Use?

- **Meta-evaluation** runs human evaluation and automatic evaluation on the same outputs, calculates correlation
- Examples:
  - **WMT Metrics Task** for MT (Mathur et al. 2021)
  - **RealSumm** for summarization (Bhandari et al. 2020)
- Evaluation is hard, especially with good systems!
  Most metrics had no correlation w/ human eval over best systems

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).

Class discussion

- Pick a 4-line excerpt from a short text (e.g. poem, text message) in English

- Use Google translate, VolcTrans(translate.volcengine.com), ChatGPT to back-translate the text via a pivot language, e.g.,
  - English → Spanish → English
  - English → L1 → L2 → English, where L1 and L2 are typologically different from English and from each other

- Compare the original text and its English back-translation, and share your observations. For example,
  - What information got lost in the process of translation?
  - Are there translation errors associated with linguistic properties of pivot languages and with linguistic divergences across languages?
  - Try different pivot languages: can you provide insights about the quality of MT for those language pairs?