Cracking Multi-Word Expressions *in a nutshell:* Neural MT walks on thin ice when facing MWEs

The University of Manchester (*current*), UK & ADAPT Research Centre, DCU (*former*), Ireland

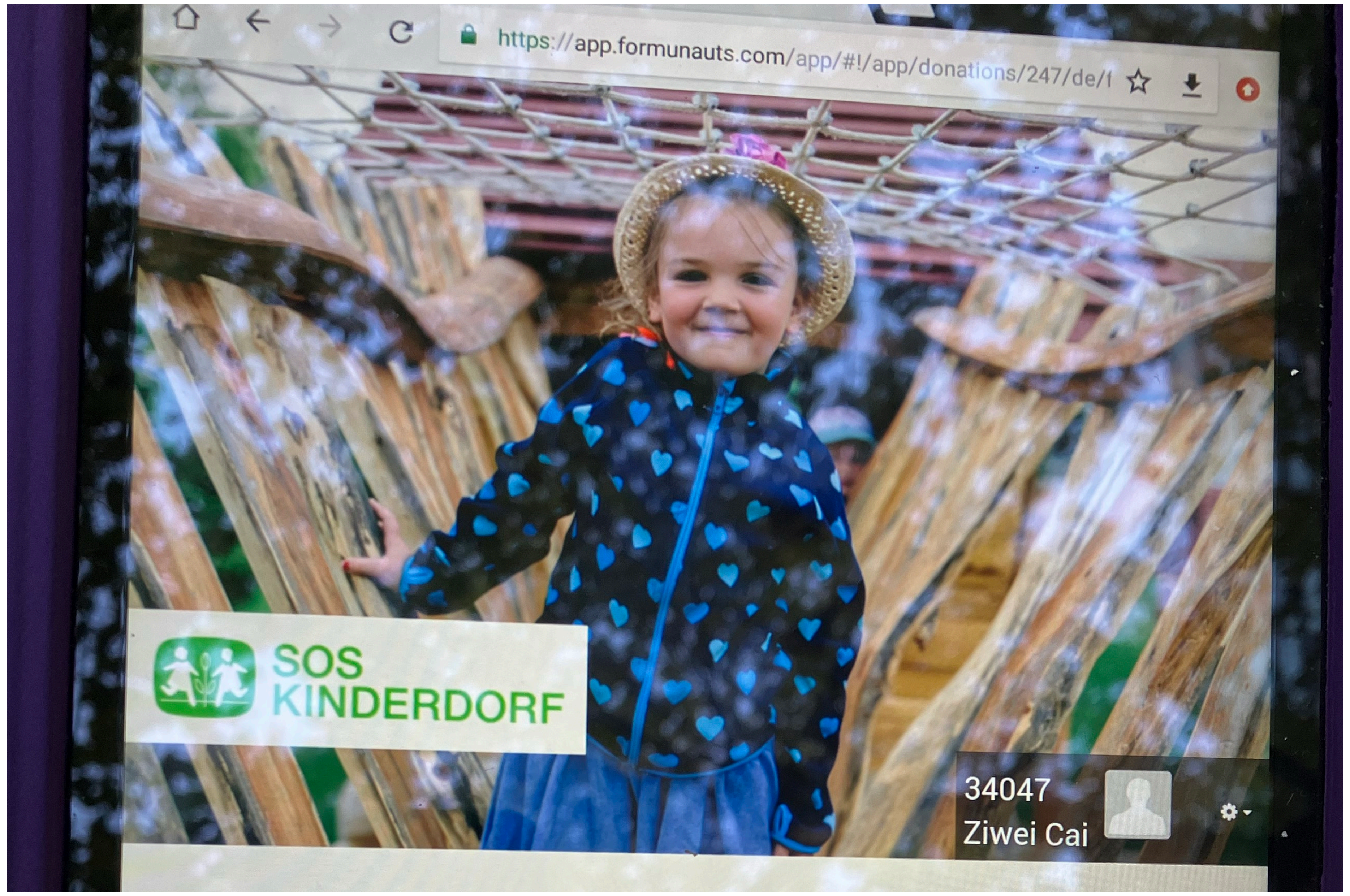
Lifeng Han (Aaron)

lifeng.Han@manchester.ac.uk

Oct. 11th, 2022 @ Institute for Natural Language Processing (IMS) Uni.Stuttgart


Meta-Evaluation of Translation Evaluation Methods:
a systematic up-to-date overview
Lifeng Han * and Serge Gladkoff **
* The University of Manchester (current), UK & ADAPT Research Centre, DCU (former), Ireland
** Logrus Global (https://logrusglobal.com)
Half-day Tutorial @ LREC2022, June 20th, Marseille, France

LREC22-trip



My first experience in stuttgart

Their desk in city centre

University of Stuttgart

Website | Directions | Save | Call

University in Stuttgart, Germany

The University of Stuttgart is a leading research university located in Stuttgart, Germany. It was founded in 1829 and is organized into 10 faculties. It is one of the oldest technical universities in Germany with highly ranked programs in civil, mechanical, industrial and electrical engineering, among others. Wikipedia

Address: Keplerstraße 7, 70174 Stuttgart, Germany

Phone: +49 711 6850

Undergraduate tuition and fees: Domestic tuition 650 EUR, International tuition 1,500 EUR (2019 – 20)

Total enrollment: 25,705 (2018)

Founder: William I of Württemberg

Founded: 1829

Rector: Wolfram Ressel

History

Graduation rate

Philosophy

Via Google KG - sister college to UoM? (Why not 'brother college'? https://en.wikipedia.org/wiki/Sister_college)

Abstract

In this talk, I will present some work related to multi-word expressions (MWEs), including MWE identification and translation. The presentation covers

1) our participation on **MWE identification** back in MWE2017@EACL,

2) our neural machine translation (NMT) models addressing **MWE translation from low-frequency terms** perspective for EN-DE/ZH, which includes bilingual MWE terms extraction and augmentation to training data, as well as decomposing Chinese characters/symbols into lower level representations, and

3) our **multilingual parallel corpus** preparation using PARSEME English seed corpus that has verbalMWE annotations, into DE/ZH/PL and other ongoing languages, where we give many examples when NMT went wrong in the task of translating MWE related content and MWE terms, and we try to classify these errors into different types to facilitate further research.

In the end of the talk, you will realise that when MT has reached a new level of quality via NMT, MWEs become very apparent bottlenecks in front of MT researchers.

We also propose some possible solutions to address these issues.

Some resources used by this talk were included in

- LREC22 tutorial "Meta-Evaluation of Translation Evaluation Methods: a systematic up-to-date overview" Han and Gladkoff https://github.com/poethan/LREC22_MetaEval_Tutorial
- PhD thesis (2022), "An investigation into multi-word expressions in machine translation" ADAPT centre/school of Computing, DCU, Ireland
- NoDaLiDa21: "Chinese Character Decomposition for Neural MT with Multi-Word Expressions"
- LREC2020: "MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora" Data/Git: https://github.com/poethan/MWE4MT
- MWE2020: "AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations"
- MWE2017: "Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking" Maldonado, Han, Moreau, et al.

Abbreviations/acronyms

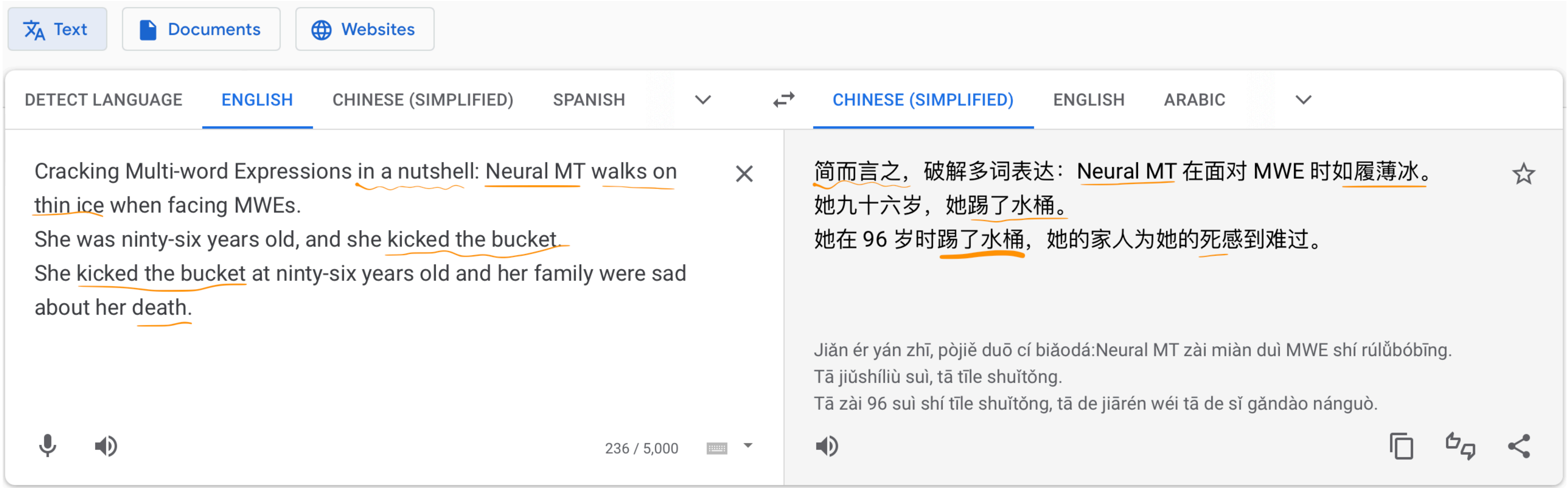
- MWEs: multi-word expressions
- MT: machine translation
- SMT: statistical machine translation
- NMT: neural machine translation
- MTE: machine translation evaluation
- TQA: translation quality assessment
- TQE: translation quality evaluation
- QE: quality estimation
- HumanEval (HE): human evaluation
- AutoEval (AE): automatic evaluation
- MetaEval: meta-evaluation (evaluating the evaluation methods)
- NE: named entities
- EN/PL/DE/ZH: English/Polish/German/Chinese
- CRFs: Conditional Random Fields

On the title

in a nutshell (from dictionary)
in the fewest possible words: *she put the matter in a nutshell*.
[originally with allusion to a copy of Homer's Iliad which, as reported by Pliny the Elder, was supposedly small enough to be enclosed in the shell of a nut.]

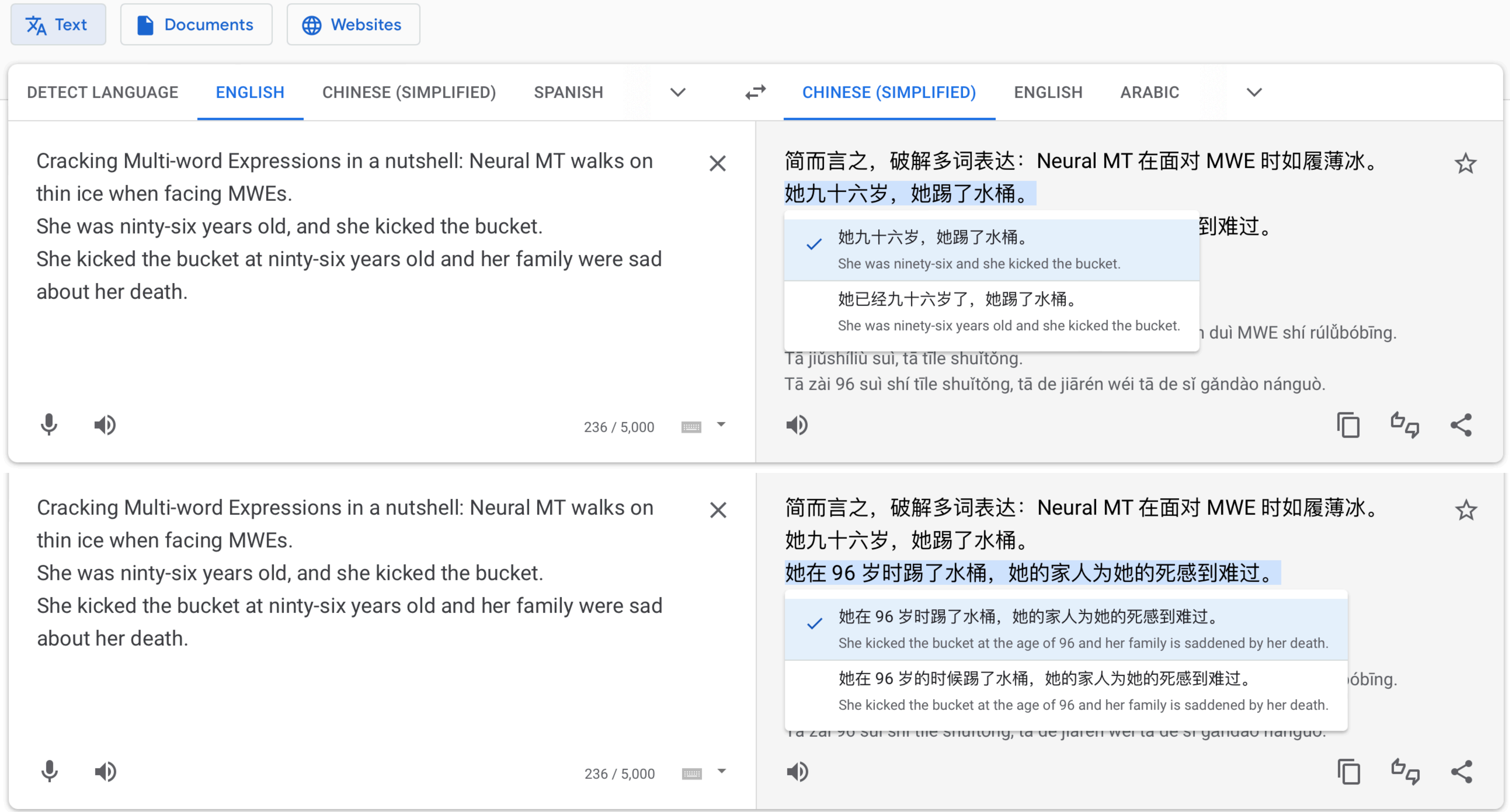
On the title

Un/intentional correct? Frequency?



googleMT: 2022.10.05th
Remove the full stop from the end of the first sentence, check the difference, and why? Same for the second/thrid sentence.

Either statistically enhanced, or attention-aligned, monolingual-paraphrase -> NMT?
Or Translation Memories (TMs) to NMT? Semantic re-ranking of NMT outputs?



No alignment/replacement from ‘kick the bucket’ to ‘death’ in en-zh NMT

He kicked the bucket.
kick the bucket
kicks the bucket
kicked the bucket
They kick the bucket.
Did he kick the bucket?
He was very sick and he kicked the bucket.
He kicked the bucket and all his families are sad.
Does kicking the bucket mean someone dies?
He kicked the bucket and the doctors are so sad that they could not save his life.

GoogleNMT:
So, this pair exists in en-zh corpus they used;
but why the context information is not enough to translate it correctly?
e.g. the last sentence?
How to better including context?

339 / 5,000

他踢了水桶。
氣絕
踢水桶
踢了桶
他們踢水桶。
他踢水桶了嗎？
他病得很重，他踢了水桶。
他踢了水桶，他所有的家人都很傷心。
踢水桶意味著有人死嗎？
他踢了水桶，醫生很傷心，他們無法挽救他的生命。

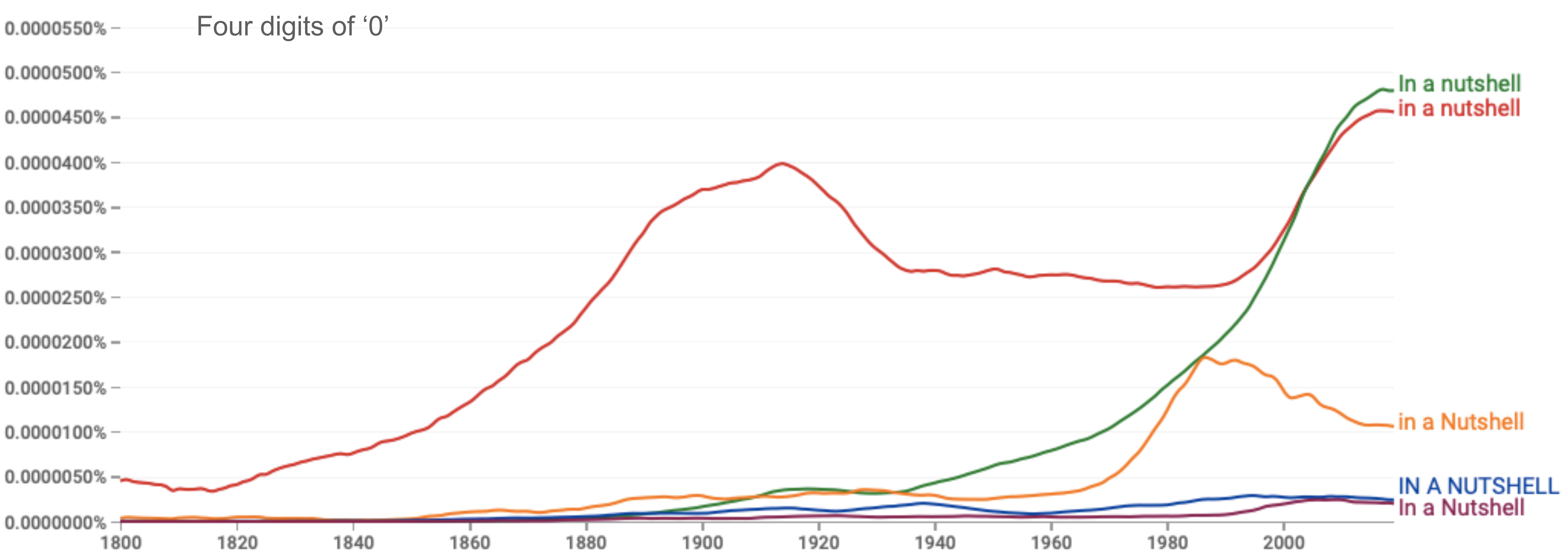
The only one saying “death”

Un-intentional correct?

Google Books Ngram Viewer

in a nutshell

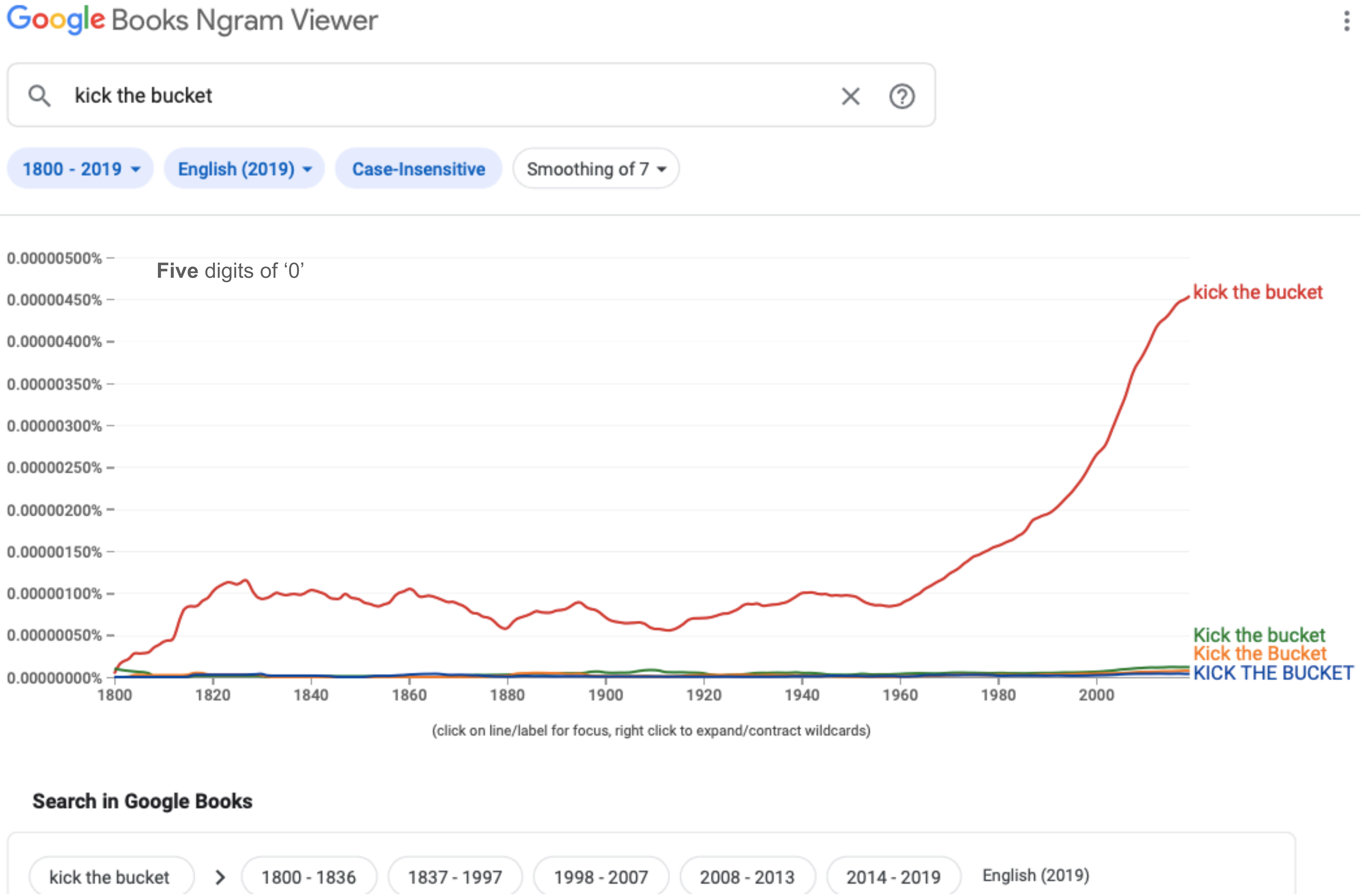
1800 - 2019 | English (2019) | Case-Insensitive | Smoothing of 7



(click on line/label for focus, right click to expand/contract wildcards)

Search in Google Books

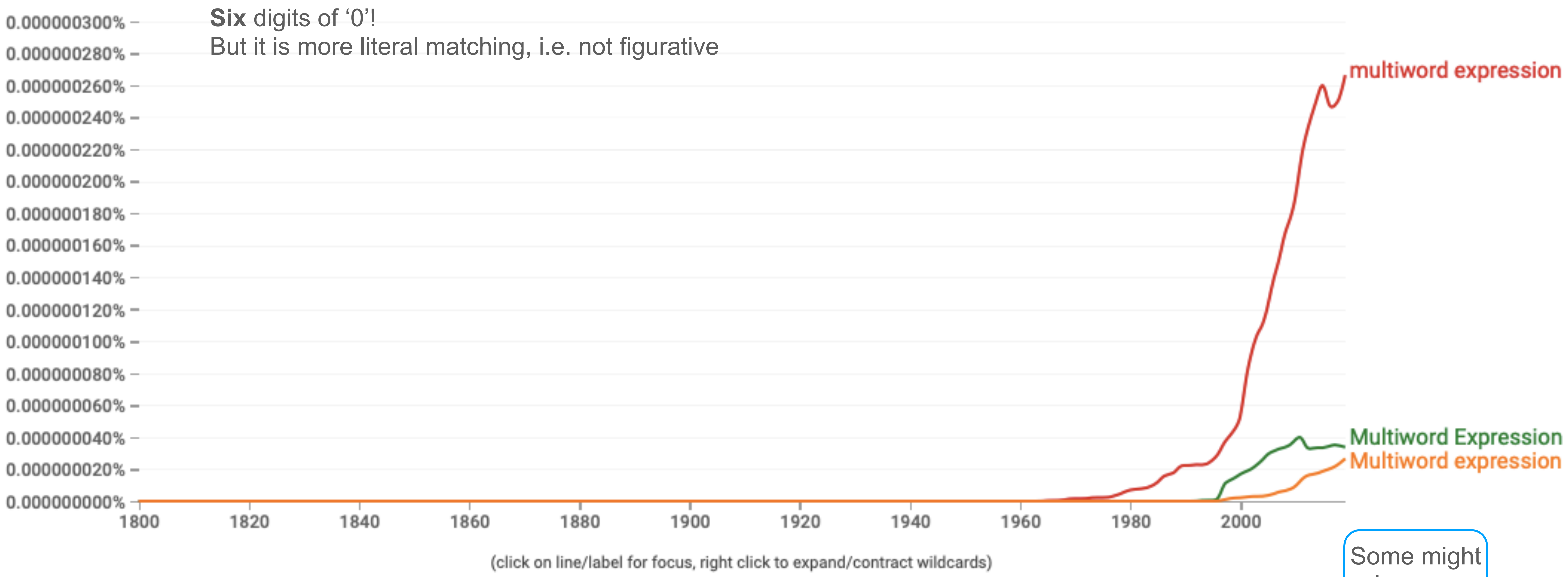
in a nutshell > 1800 - 1867 | 1868 - 1904 | 1905 - 1917 | 1918 - 2007 | 2008 - 2019 | English (2019)


Google Books Ngram Viewer
kick the bucket
1800 - 2019
English (2019)
Case-Insensitive
Smoothing of 7
Five digits of ‘0’
0.00000500%
0.00000450%
0.00000400%
0.00000350%
0.00000300%
0.00000250%
0.00000200%
0.00000150%
0.00000100%
0.00000050%
0.00000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
kick the bucket
Kick the bucket
Kick the Bucket
KICK THE BUCKET
(click on line/label for focus, right click to expand/contract wildcards)
Search in Google Books
kick the bucket
1800 - 1836
1837 - 1997
1998 - 2007
2008 - 2013
2014 - 2019
English (2019)

Google Books Ngram Viewer

multiword expression

1800 - 2019 | English (2019) | Case-Insensitive | Smoothing of 7



Some might be your work!

Search in Google Books

multiword expression > 1800 - 2000 | 2001 - 2016 | 2017 | 2018 | 2019 | English (2019)

walk on thin ice

1. To act or proceed with great care, caution, and consideration so as not to upset someone or trigger some imminent disaster.

The littlest thing tends to anger my mother, so I feel like I have to walk on thin ice whenever I'm at her house.

I suggest walking on thin ice if you plan to go after those lobbyists—they have the power and influence to do some serious damage to your career.

2. To do something very risky or dangerous that could result in imminent repercussions or disaster.

Tim flaunts the way he walks on thin ice with his risky investments, but one of these days he's going to end up losing it all.

You're walking on thin ice by continuing to come in late like that. If the boss notices, you'll be fired for sure.

See also: **ice**, **on**, **thin**, **walk**

walk on thin ice *verb*

See **walk on eggs**

See also: **ice**, **on**, **thin**, **walk**

walk on eggs

To act or proceed with great care, caution, and consideration so as not to upset someone or trigger some imminent disaster.

The littlest thing tends to anger my mother, so I feel like I have to walk on eggs whenever I'm at her house.

You always have to walk on eggs when you talk to the media about the company, because they can end your career in a heartbeat if they feel like you're exposing their business practices to public scrutiny.

See also: **egg**, **on**, **walk**

walking on eggs

Acting or proceeding with great care, caution, and consideration so as not to upset someone or trigger some imminent disaster.

I suggest walking on eggs if you plan to go after those lobbyists—they have the power and influence to do some serious damage to your career.

The littlest thing tends to anger my mother, so I feel like I am walking on eggs whenever I'm at her house.

See also: **egg**, **on**, **walking**

Google Translate

Text Documents Websites

GUAGE CHINESE (TRADITIONAL) CHINESE (SIMPLIFIED) ENGLISH

CHINESE (TRADITIONAL) ENGLISH CHINESE (SIMPLIFIED)

walk on thin ice
walk on eggs
To act or proceed with great care, caution, and consideration so as not to upset someone or trigger some imminent disaster.
The littlest thing tends to anger my mother, so I feel like I have to walk on eggs whenever I'm at her house.
You always have to walk on eggs when you talk to the media about the company, because they can end your career in a heartbeat if they feel like you're exposing their business practices to public scrutiny.

469 / 5,000

在薄冰上行走
走在雞蛋上
以非常小心、謹慎和考慮的方式行事或進行，以免使某人不安或引發一些迫在眉睫的災難。
最小的事情往往會激怒我的母親，所以每當我在她家時，我都覺得我必須在雞蛋上行走。
當你在媒體上談論公司時，你總是必須腳踏實地，因為如果他們覺得你將他們的商業行為暴露在公眾監督之下，他們可能會在心跳中結束你的職業生涯。

During the heartbeat

Zài bó bīng shàng xíngzǒu
zǒu zài jīdàn shàng
yǐ fēicháng xiǎoxīn, jǐnshèn hé kǎolǜ de fāngshì xíngshì huò jìnxíng, yǐmiǎn shǐ mǒu rén

Show more

On MWE: definitions and scopes

- Baldwin and Kim (2010) define MWEs as “lexical items that: (i) can be decomposed into multiple lexemes; and (ii) display lexical, syntactic, semantic, pragmatic or statistical idiomaticity”.
- MWEs have a broad coverage/scope of linguistic phenomena: syntax, semantics, and pragmatics.
- - E.g. compound words (nouns), named entities, verbal MWEs (verb particle constructions, light verb constructions), discourse markers, collocations, lexical bundles, idioms and metaphors, and more.
- MWEs can appear in unexpected syntax and can lead to ambiguity regarding whole sentence understanding.

- Timothy Baldwin and Su Nam Kim. “Multiword Expressions”. In: Hand- book of Natural Language Processing, Second Edition. Chapman and Hall, 2010, pp. 267–292.

More examples on scopes

- In English, verb-noun patterns ``jump the gun'' and ``kick the bucket'', noun-noun combinations ``hit man'' and ``word salad”, adjective-noun combinations like ``cold shoulder'' and ``hot dog'', binomial expressions such as ``by and large'' and ``nuts and bolts’' (Socolof et al. 2021; Han 2022 thesis).
- MWEs can be {continuous} where all component words form an n-gram together without gap words vs {dis-continuous} with some other words inserted in-between.
- - e.g. pick … up, ``去 (qù) ... 了 (le)" meaning ``went to ...(do something, or somewhere)”, “唯...是瞻 (wéi … shì zhān)” (only look at …, 唯余马首是瞻)
- MWE has deep implications for NLP systems: MT, NLG, Parsing, Lexicography, Computer-aided language learning (CALL)

Traffic jams, Brain storming, 诗与远方(shī yǔ yuǎnfāng, poem and distance)

On MWE discovery and identification

Shared tasks and our models

- SIGLEX-MWE: special interest group in lexicon - MWE section, ACL
- MWE-workshops: 18th in MWE2022
- “MWE discovery and identification shared tasks” with MWE WS, from 2017
- - supported by PARSEME project
- - monolingual tasks, covering mostly European languages.
- KONVENS21: benchmark on German verbal idiom disambiguation.

On MWE discovery and identification

Sub-sentential multi-word constructions that involves a verb

- ID: Idiomatic/fixed expressions (oft. Non-compositional)
  - let the cat out of the bag – kick the bucket
  - spill the beans
- LVC: Light Verb Constructions
  - make a decision – take a photo
  - have a convesation – have courage
- VPC: Verb-Particle Constructions (phrasal verbs)
  - take off – look up [something] / look [something] up
  - put up – ask [someone] out
- IReflV: Reflexives (in French, Spanish, etc.)
  - se dérouler (‘to unfold’) – se passer (‘to happen’)
  - se trouver (‘to be located’) – se battre (‘to fight’, ‘to strive’)
- OTH: Other, potentially language-specific, VMWEs
  - drink and drive – pretty-print – tumbly dry

On MWE discovery and identification

18 languages in total

– (BG) Bulgarian	(CS) Czech	(DE) German
– (EL) Greek	(ES) Spanish	(FA) Farsi
– (FR) French	(HE) Hebrew	(IT) Italian
– (LT) Lithuanian	(MT) Maltese	(PL) Polish
– (PT) Br. Portuguese	(RO) Romanian	(SL) Slovenian
– (SV) Swedish	(TR) Turkish	

15 provided with morpho-syntactic information (POS, lemmas and/or dependency info)

We did not attempt 3 langs with no morpho-syntactic info (BG, HE, LT)

On MWE discovery and identification

- Two tracks:
  - Closed Track – Systems permitted to only use provided data
  - Open Track – You're free to use provided data and/or any other external data (MWE lexicons, symbolic rules, wordnets, raw corpora, word embeddings, language models trained on external data, etc.)
- We were the only team that (sort of) attempted both tracks

On MWE discovery and identification

- But you're not required to categorise your extracted VMWEs (binary evaluation)
- MWE-based vs Token-based evaluation
- There can be non-contiguous VMWEs
- One sentence can have 0, 1 or more VMWEs
- VMWEs can be embedded / overlapped
- VMWEs can become a single-word token (e.g. relfexives in Spanish: *encontrarse* vs *se encuentra*)

On MWE discovery and identification

PARSEME TSV Format

```
# sentid: Europar.550_00129
# sentence-text: Je voudrais seulement rappeler qu'il se passe
  actuellement quelque chose de terrible en Serbie.
1     Je                   _     _
2     voudrais             _     _
3     seulement            _     _
4     rappeler             _     _
5     qu'                  nsp   _
6     il                   _     _
7     se                   _     1:IReflV
8     passe                _     1
9     actuellement         _     _
10    quelque              _     _
11    chose                _     _
12    de                   _     _
13    terrible             _     _
14    en                   _     _
15    Serbie               nsp   _
16    .                    _     _
```

Sentence means:
I would only like to remind (you) that there is something terrible happening in Serbia right now.

On MWE discovery and identification

CONLLU TSV Format

sentid: Europar.550_00129
sentence-text: Je voudrais seulement rappeler qu'il se passe actuellement quelque chose de terrible en Serbie.

1	Je	il	CL	CLS	n=s|p=1|s=suj|sentid=Europar.550_00129	2	suj	_
2	voudrais	vouloir	V	V	m=ind|n=s|p=1|t=cond	0	root	_
3	seulement	seulement	ADV	ADV	_	2	mod	_
4	rappeler	rappeler	V	VINF	m=inf	2	obj	_
5	qu'	que	C	CS	s=s	4	obj	_
6	il	il	CL	CLS	g=m|n=s|p=3|s=suj	8	suj	_
7	se	le/lui	CL	CLR	p=3|s=refl	8	aff	_
8	passe	passer	V	V	m=ind|n=s|p=3|t=pst	5	obj.cpl	_
9	actuellement	actuellement	ADV	A	DV	8	mod	_
10	quelque	quelque	D	DET	mwehead=PRO|mwelemma=quelque_chose...	8	obj	_
11	chose	chose	N	NC	component=y|g=f|n=s|s=c	10	dep_cpd	_
12	de	de	P	P	_	10	dep	_
13	terrible	terrible	A	ADJ	n=s|s=qual	12	obj.p	_
14	en	en	P	P	_	8	mod	_
15	Serbie	Serbie	N	NPP	g=f|n=s|s=p	14	obj.p	_
16	.	.	PONCT	PONCT	s=s	2	ponct	_

Syntactic dependencies!!!

On MWE discovery and identification

Expected output:

- Same format as input
- Give numeric ID to identified VMWE per sentence

Evaluation:

- F1 measure
- MWE-based evaluation
- Token-based (unigram-based) evaluation

On MWE discovery and identification

- Intuition 1: Expected output format suggests a token tagging problem
  - Like Named-Entity Recognition or POS tagging
- Approach 1 (**CRF**): Use a sequence model to predict VMWEs

Intuition 2: The context of a word will be different when used in an idiomatic expression than when used on its own.

- "the other side of the coin" → sentences about any topic contrasting ideas or situations
- "coin" **not** occurring in "the other side of the coin" → sentences about currency, numismatics, etc.
- Intuition exploited by word-sense disambiguation

Approach 2 (**Semantic Reranker**): Given a VMWE candidate, build vectors of its content words using a large corpus. Compute pairwise semantic similarity comparisons (e.g. cosine scores) (Maldonado and Emms, 2011)

[Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking]
Alfredo Maldonado, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel and Qun Liu
The 13th Workshop on Multiword Expressions (MWE 2017) @ EACL2017.

On MWE discovery and identification

Approach 1 (CRF) output qualifies for Closed Track

Approach 2 (Semantic Reranker) output would qualify for Open Track

Our best CRF prediction directly submitted as our Closed Track participation

10 best CRF predictions reranked by Approach 2 – best reranked prediction would be our Open Track participation

Unfortunately the semantic reranker was not ready on time for the competition

On MWE discovery and identification

Settled on **CRF++** (https://taku910.github.io/crfpp/)

Built conversion scripts to **merge** CONLLU and PARSEME TSV features

Designed a few CRF++ **feature templates**

Conducted **5-fold cross validation** experiments on **Training Set** using feature templates

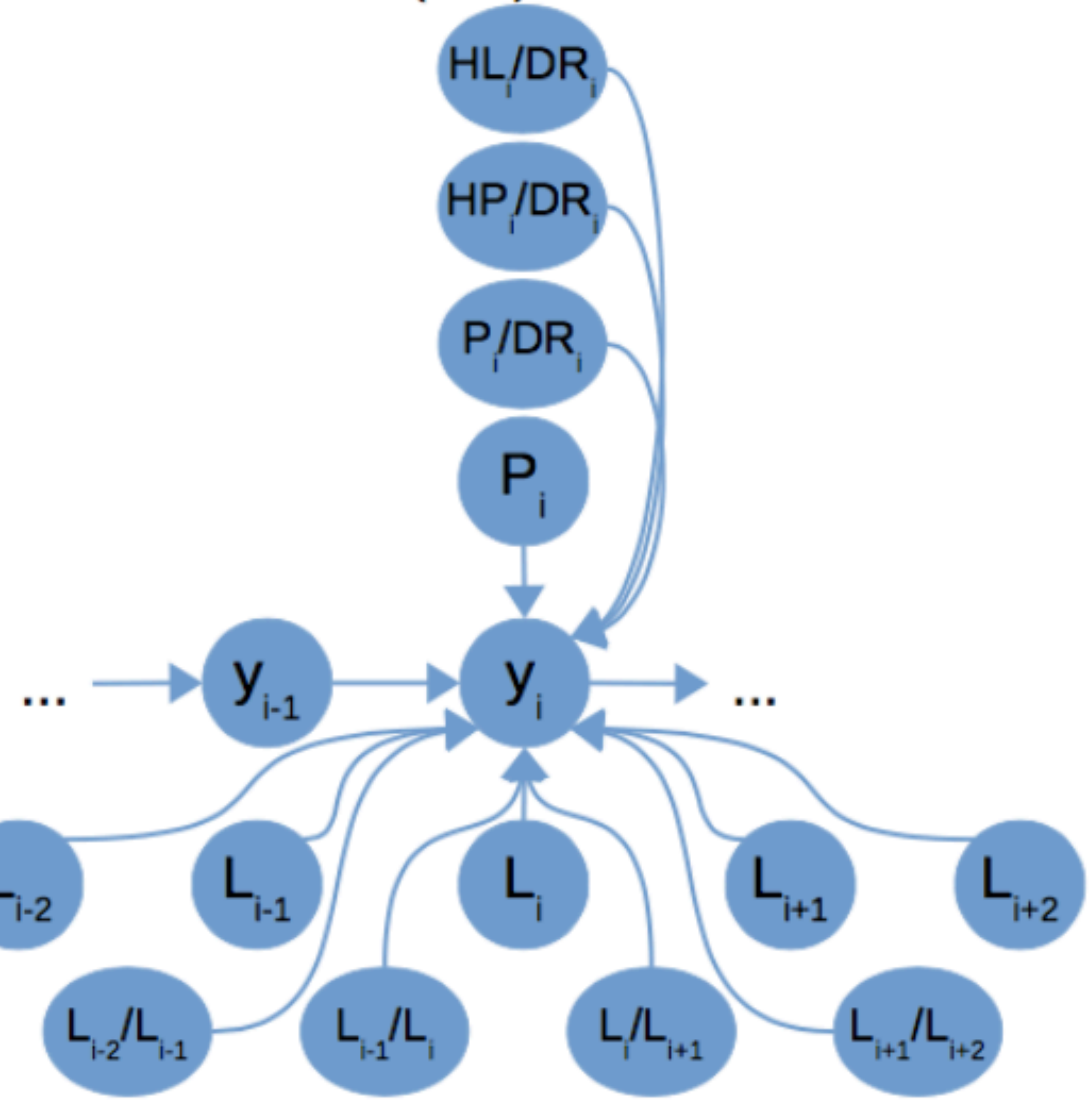
Best performing template per language used to train on **full Training Set**

Best performing template per language used to **Predict** on **Blind Test Set**

On MWE discovery and identification

For each token we have:

- Word surface (W) – Lemma (L) – POS (P)
- Head word surface (HW) – Head lemma (HL) – Head POS
- Dependency relation with head (DR)



On MWE discovery and identification

Ash.:

Europarl corpus (Koehn, 2005) for building word vectors

- Not available for Farsi, Maltese, Turkish

Developed code to build word vectors from candidate MWEs in each of the **top 10 CRF predicted sequences**

MWEs' word vectors are combined to produce **1 vector representing 1 CRF predicted sequence**

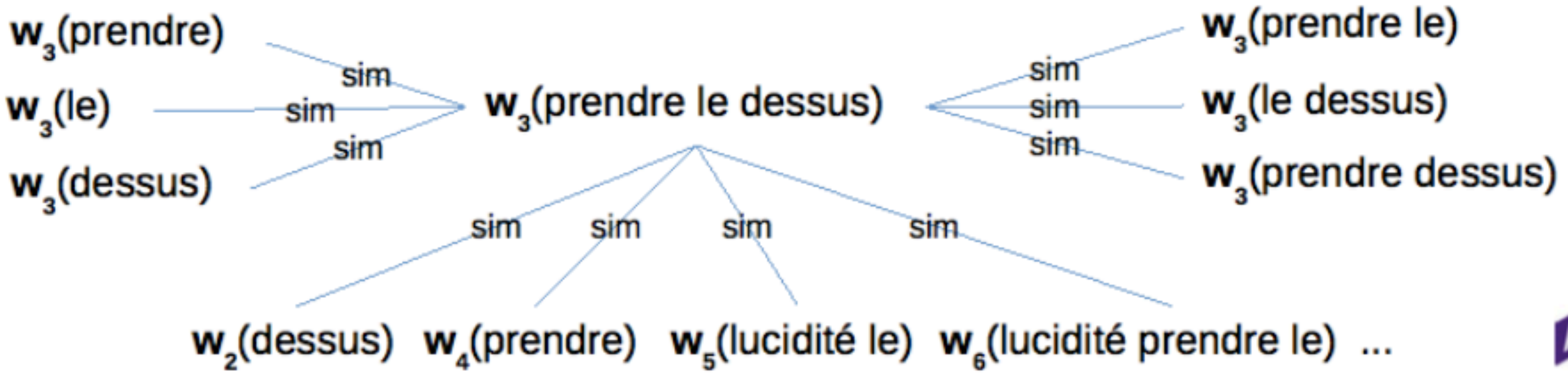
Sequence vectors fed to a **decision-tree regression** algorithm from Weka (Hall et al., 2009)

Regression produces a score in [0, 1] per predicted sequence – Score used to rerank predictions

On MWE discovery and identification

id	W	P	L	HL	HP	DR	Gold	Best 10 sequences 1	2	3	4	5	6	7	8	9	10
1	L'	DET	le	écarteur	NOUN	det	_	_	_	_	_	_	_	_	_	_	_
2	écarteur	NOUN	écarteur	conscient	ADJ	nsubj	_	_	_	_	_	_	_	_	_	_	_
3	est	VERB	être	conscient	ADJ	cop	_	_	_	_	_	_	_	_	_	_	_
4	parfaitement	ADV	parfaitement	conscient	ADJ	advmod	_	_	_	_	_	_	_	_	_	_	_
5	conscient	ADJ	conscient	<[ROOT]>	<[ROOT]>	root	_	_	_	_	_	_	_	_	_	_	_
6	de	ADP	de	état	NOUN	case	_	_	_	_	_	_	_	_	_	_	_
7	son	DET	son	état	NOUN	nmod:poss	_	_	_	_	_	_	_	_	_	_	_
8	état	NOUN	état	conscient	ADJ	nmod	_	_	_	_	_	_	_	_	_	_	_
9	dans	ADP	dans	moment	NOUN	case	_	_	_	_	_	_	_	_	_	_	_
10	les	DET	le	moment	NOUN	det	_	_	_	_	_	_	_	_	_	_	_
11	moments	NOUN	moment	conscient	ADJ	nmod	_	_	_	_	_	_	_	_	_	_	_
12	où	PRON	où	prendre	VERB	nmod	_	_	_	_	_	_	_	_	_	_	_
13	les	DET	le	douleur	NOUN	det	_	_	_	_	_	_	_	_	_	_	_
14	douleurs	NOUN	douleur	prendre	VERB	nsubj	_	_	_	_	_	_	_	_	_	_	_
15	et	CONJ	et	douleur	NOUN	cc	_	_	_	_	_	_	_	_	_	_	_
16	où	PRON	où	prendre	VERB	nmod	_	_	_	_	_	_	_	_	_	_	_
17	sa	DET	son	lucidité	NOUN	nmod:poss	_	_	_	_	_	_	_	_	_	_	_
18	lucidité	NOUN	lucidité	douleur	NOUN	conj	_	_	_	_	_	B	_	_	_	I	_
19	prennent	VERB	prendre	moment	NOUN	acl:relcl	B	_	_	B	B	_	I	_	I	_	B
20	le	DET	le	dessus	NOUN	det	I	_	_	I	_	I	I	I	_	I	I
21	dessus	NOUN	dessus	prendre	VERB	dobj	I	_	I	I	_	_	I	I	_	_	_
22	.	PUNCT	.	conscient	ADJ	punct	_	_	_	_	_	_	_	_	_	_	_

- Choice of Similarity scores:
  - Cosine w IDF
  - Cosine w/o IDF
  - Jaccard
  - Min/Max Sim
- Summarised with mean, min & max to produce seq. vector



On MWE discovery and identification

								Best 10 sequences									
id	W	P	L	HL	HP	DR	Gold	1	2	3	4	5	6	7	8	9	10
1	L'	DET	le	écarteur	NOUN	det	_	_	_	_	_	_	_	_	_	_	_
2	écarteur	NOUN	écarteur	conscient	ADJ	nsubj	_	_	_	_	_	_	_	_	_	_	_
3	est	VERB	être	conscient	ADJ	cop	_	_	_	_	_	_	_	_	_	_	_
4	parfaitement	ADV	parfaitement	conscient	ADJ	advmod	_	_	_	_	_	_	_	_	_	_	_
5	conscient	ADJ	conscient	<[ROOT]>	<[ROOT]>	root	_	_	_	_	_	_	_	_	_	_	_
6	de	ADP	de	état	NOUN	case	_	_	_	_	_	_	_	_	_	_	_
7	son	DET	son	état	NOUN	nmod:poss	_	_	_	_	_	_	_	_	_	_	_
8	état	NOUN	état	conscient	ADJ	nmod	_	_	_	_	_	_	_	_	_	_	_
9	dans	ADP	dans	moment	NOUN	case	_	_	_	_	_	_	_	_	_	_	_
10	les	DET	le	moment	NOUN	det	_	_	_	_	_	_	_	_	_	_	_
11	moments	NOUN	moment	conscient	ADJ	nmod	_	_	_	_	_	_	_	_	_	_	_
12	où	PRON	où	prendre	VERB	nmod	_	_	_	_	_	_	_	_	_	_	_
13	les	DET	le	douleur	NOUN	det	_	_	_	_	_	_	_	_	_	_	_
14	douleurs	NOUN	douleur	prendre	VERB	nsubj	_	_	_	_	_	_	_	_	_	_	_
15	et	CONJ	et	douleur	NOUN	cc	_	_	_	_	_	_	_	_	_	_	_
16	où	PRON	où	prendre	VERB	nmod	_	_	_	_	_	_	_	_	_	_	_
17	sa	DET	son	lucidité	NOUN	nmod:poss	_	_	_	_	_	_	_	_	_	_	_
18	lucidité	NOUN	lucidité	douleur	NOUN	conj	_	_	_	_	_	B	_	_	_	I	_
19	prennent	VERB	prendre	moment	NOUN	acl:relcl	B	_	_	B	B	_	I	_	I	_	B
20	le	DET	le	dessus	NOUN	det	I	_	_	I	_	I	I	I	_	I	I
21	dessus	NOUN	dessus	prendre	VERB	dobj	I	_	I	I	_	_	I	I	_	_	_
22	.	PUNCT	.	conscient	ADJ	punct	_	_	_	_	_	_	_	_	_	_	_
							CRF	.3919	.3861	.2033	.0066	.0065	.0019	.0012	.0006	.0006	.0004
							Sem	.3589	.0113	.5011	.0216	.0216	.0216	.0216	.0325	.0325	.0325

CS

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.7897	0.6560	0.7167	1	0.8246	0.6655	0.7365	1
ADAPT	closed	0.5931	0.5621	0.5772	3	0.8191	0.6561	0.7286	2
RACAI	closed	0.7009	0.5918	0.6418	2	0.8190	0.6228	0.7076	3
MUMULS	closed	0.4413	0.1028	0.1667	4	0.7747	0.1387	0.2352	4

DE

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
SZEGED	closed	0.5154	0.3340	0.4053	2	0.6592	0.3468	0.4545	1
TRANSITION	closed	0.5503	0.3280	0.4110	1	0.5966	0.3133	0.4109	2
ADAPT	closed	0.3308	0.1740	0.2280	3	0.7059	0.2837	0.4048	3
MUMULS	closed	0.3277	0.1560	0.2114	4	0.6988	0.2286	0.3445	4
RACAI	closed	0.3652	0.1300	0.1917	5	0.6716	0.1793	0.2830	5

EL

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.3612	0.4500	0.4007	1	0.4635	0.4742	0.4688	1
ADAPT	closed	0.3437	0.2880	0.3134	4	0.5380	0.3601	0.4314	2
MUMULS	closed	0.2087	0.2580	0.2308	5	0.4294	0.4143	0.4217	3
SZEGED	closed	0.3084	0.3300	0.3188	2	0.4451	0.3757	0.4075	4
RACAI	closed	0.4286	0.2520	0.3174	3	0.5616	0.2953	0.3871	5

ES

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.6122	0.5400	0.5739	1	0.6574	0.5252	0.5839	1
ADAPT	closed	0.6105	0.3480	0.4433	2	0.7448	0.3670	0.4917	2
MUMULS	closed	0.3673	0.3100	0.3362	4	0.6252	0.3995	0.4875	3
SZEGED	closed	0.2575	0.5000	0.3399	3	0.3635	0.5629	0.4418	4
RACAI	closed	0.6447	0.1960	0.3006	5	0.7233	0.1967	0.3093	5

FA

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.8770	0.8560	0.8664	1	0.9159	0.8885	0.9020	1
ADAPT	closed	0.7976	0.8040	0.8008	2	0.8660	0.8416	0.8536	2

FR

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
ADAPT	closed	0.6147	0.4340	0.5088	2	0.8088	0.4964	0.6152	1
TRANSITION	closed	0.7484	0.4700	0.5774	1	0.7947	0.4856	0.6028	2
RACAI	closed	0.7415	0.3500	0.4755	3	0.7872	0.3673	0.5009	3
SZEGED	closed	0.0639	0.0520	0.0573	6	0.5218	0.2482	0.3364	4
MUMULS	closed	0.1466	0.0680	0.0929	5	0.5089	0.2067	0.2940	5
LIF	closed	0.8056	0.0580	0.1082	4	0.8194	0.0532	0.1000	6
LATL	open	0.4815	0.4680	0.4746	1	0.5865	0.5108	0.5461	1

HU

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
SZEGED	closed	0.7936	0.6934	0.7401	1	0.8057	0.6317	0.7081	1
MUMULS	closed	0.6291	0.6152	0.6221	5	0.7132	0.6657	0.6886	2
TRANSITION	closed	0.6484	0.7575	0.6987	2	0.6502	0.7012	0.6747	3
ADAPT	closed	0.7570	0.5992	0.6689	3	0.7846	0.5710	0.6610	4
RACAI	closed	0.8029	0.5471	0.6508	4	0.8208	0.5015	0.6226	5

IT

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.5354	0.3180	0.3990	1	0.6134	0.3378	0.4357	1
SZEGED	closed	0.1503	0.1560	0.1531	4	0.4054	0.3064	0.3490	2
ADAPT	closed	0.6174	0.1420	0.2309	2	0.6964	0.1532	0.2511	3
RACAI	closed	0.6125	0.0980	0.1690	3	0.6837	0.1053	0.1824	4

MT

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.1565	0.1340	0.1444	1	0.1843	0.1460	0.1629	1
ADAPT	closed	0.2043	0.0380	0.0641	2	0.3084	0.0518	0.0887	2
RACAI	closed	0.2333	0.0280	0.0500	3	0.2481	0.0259	0.0469	3
MUMULS	closed	0.0000	0.0000	0.0000	n/a	0.0000	0.0000	0.0000	n/a

PL

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
ADAPT	closed	0.7798	0.6020	0.6795	2	0.8742	0.6228	0.7274	1
TRANSITION	closed	0.7709	0.6260	0.6909	1	0.8000	0.6312	0.7056	2
MUMULS	closed	0.6562	0.5460	0.5961	3	0.8310	0.6013	0.6977	3
SZEGED	closed	0.0000	0.0000	0.0000	n/a	0.0000	0.0000	0.0000	n/a

PT

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.7543	0.6080	0.6733	1	0.8005	0.6370	0.7094	1
ADAPT	closed	0.6410	0.5320	0.5814	2	0.8348	0.6054	0.7018	2
MUMULS	closed	0.5358	0.3740	0.4405	3	0.8247	0.4717	0.6001	3
SZEGED	closed	0.0129	0.0080	0.0099	4	0.6837	0.1987	0.3079	4

RO

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
MUMULS	closed	0.7683	0.7760	0.7721	2	0.8620	0.8112	0.8358	1
ADAPT	closed	0.7548	0.7140	0.7338	4	0.8832	0.7636	0.8190	2
TRANSITION	closed	0.7097	0.8020	0.7531	3	0.7440	0.8449	0.7912	3
RACAI	closed	0.8652	0.7060	0.7775	1	0.8773	0.7019	0.7799	4

SL

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.4343	0.4300	0.4322	1	0.4796	0.4522	0.4655	1
MUMULS	closed	0.3557	0.2760	0.3108	3	0.6142	0.3628	0.4562	2
ADAPT	closed	0.5142	0.2900	0.3708	2	0.7285	0.3262	0.4506	3
RACAI	closed	0.5503	0.2080	0.3019	4	0.7339	0.2145	0.3320	4

SV

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
ADAPT	closed	0.4860	0.2203	0.3032	2	0.5253	0.2249	0.3149	1
SZEGED	closed	0.2482	0.2966	0.2703	3	0.2961	0.3294	0.3119	2
TRANSITION	closed	0.5100	0.2161	0.3036	1	0.5369	0.2150	0.3070	3
RACAI	closed	0.5758	0.1610	0.2517	4	0.6538	0.1677	0.2669	4

TR

System	Track	Per-MWE results				Per-token results			
		Precision	Recall	F-measure	Rank	Precision	Recall	F-measure	Rank
TRANSITION	closed	0.6106	0.5070	0.5540	1	0.6123	0.5039	0.5528	1
ADAPT	closed	0.4541	0.4052	0.4283	3	0.5993	0.4728	0.5285	2
RACAI	closed	0.6304	0.4391	0.5176	2	0.6340	0.4348	0.5159	3
MUMULS	closed	0.4557	0.2774	0.3449	4	0.6452	0.3502	0.4540	4

1st: on FR/ PL/ PL/ SV

On MWE discovery and identification

CRF exploiting syntactic dependency features is **very competitive** at identifying VMWEs

We can rerank CRF's best predictions via semantic vector regression, **boosting performance**

Most systems perform well if there is a high proportion of Previously Seen MWEs in Test Set

Is this important? Depends…

- Machine Translation → Good performance in known MWEs desired
- Lexicography → Interest in finding **new** MWEs as opposed to **known** MWEs

F1 measure best evaluation? Depends…

- Lexicography → Recall > Precision
- CALL (few good examples) → Precision > Recall

On MWE discovery and identification

Erwan: It's important to distinguish the two main components, A and B below:

A) the unsupervised semantic similarity part, which uses Europarl to calculate "semantic features" for a sentence with expressions tagged. The goal is that these features help predict whether the tagged expressions are correct or not (note that a sentence may contain 0, 1 or several expressions). More precisely, the idea is to compute features which represent whether a candidate expression is a real MWE, by comparing frequency and semantic similarity between its individual words and the full expression. It works like this:

1) extracts all the sentences with expressions labeled from the CRF output.

2) For every expression, we build pseudo-expressions for each individual word in the expression as well as for each case of "the expression minus one word". Then for every pseudo-expression and for the full expression we compute the context vector based on Europarl, i.e. the count of every word which co-occurs with the target expression (or word) within a fixed-size window. In the features we use the frequencies for each of these pseudo-expressions, as well as the semantic similarity score between each pseudo-expression and the full expression. Originally the goal was to measure compositionality (whether the meanings of the words are combined together in the expression); but these features probably also capture how often the words appear together, which is an indication of a real expression. There is an additional set of features which consist of comparing the current expression to the other 9 candidate expressions.

3) Since we need a fixed number of features for every instance = sentence (for the supervised learning part), we must "summarize": if an expression has N words, the N values are "summarized" with the min, mean and max. Same thing for the M expressions in the sentence. In training mode we also add the probability found by the CRF as a feature.

On MWE discovery and identification

Shared tasks and our models

Erwan:

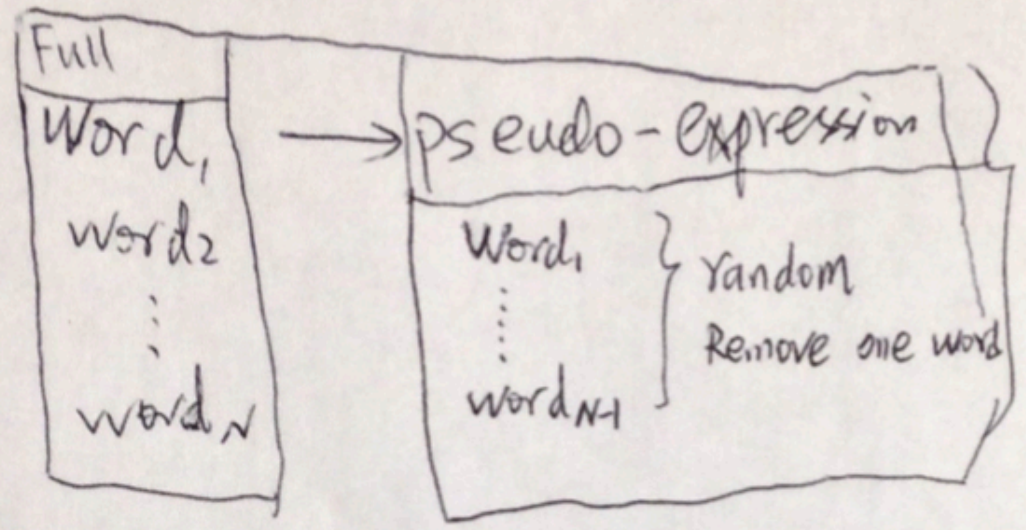
B) the supervised regression part (we used Weka decision trees regression, but other models would certainly work as well), which is fed with the features calculated using the above and predicts a single score in [0,1] which represents "how correct" the labelling of the expressions is for a sentence: here an instance is a sentence with its expressions labeled, and since for every sentence the CRF part gives us the top 10 labelling we use each of these 10 as one instance. In training mode, we assign score 1 to the gold labelling (if found among the CRF candidates) and 0 to other (wrong) labeling (the goal being to make the system assign low scores to wrong answers and high scores to good answers). In testing mode, we obtain the predicted scores and for every sentence we take the labelings which obtained the highest in the group of 10 candidates; most of the time the first from CRF is also the highest score, but sometimes the labelling we select was ranked after the first -> that's when the proper re-ranking happens.

1 Unsupervised Semantic similarity :

Using Europarl . calculate 'semantic feature' for sentence.

1.1. Extract all sentences with lables from CRF

1.2.



⇒ Context vector via Europarl

use
- frequency of pseudo-expression
- Semantic similarity score (pseudo , full)
- additional : Comparing (current expression , other 9 candidates)

1.3. Use mean score of each sentence , of different length.

⇒ Summarize

2. Supervised Regression (with weka) :

Use features from step1. calculate how correct the labeling is

During training
- Score 1 ⇒ Gold lable
- 0 ⇒ other label

Testing phase ,
- obtain predicted score for each 10 candidates
- select the highest score candidate

Abstract

In this talk, I will present some work related to multi-word expressions (MWEs), including MWE identification and translation. The presentation covers

1) our participation on **MWE identification** back in MWE2017@EACL,

2) our neural machine translation (NMT) models addressing **MWE translation from low-frequency terms** perspective for EN-DE/ZH, which includes bilingual MWE terms extraction and augmentation to training data, as well as decomposing Chinese characters/symbols into lower level representations, and

3) our **multilingual parallel corpus** preparation using PARSEME English seed corpus that has verbalMWE annotations, into DE/ZH/PL and other ongoing languages, where we give many examples when NMT went wrong in the task of translating MWE related content and MWE terms, and we try to classify these errors into different types to facilitate further research.

In the end of the talk, you will realise that when MT has reached a new level of quality via NMT, MWEs become very apparent bottlenecks in front of MT researchers.

We also propose some possible solutions to address these issues.

MT on MWEs

Looking back

- From title: on the frequency, figurative aspect of MWEs, their impact on MT
- From MWE identification: frequency matters, syntactic/semantic dependency matter.

- From low-frequency phrases/terms perspective:
- - Monolingual MWEs extraction (POS), BiMWE alignment, data augmentation
- - linguistic driven: decomposition of Chinese symbols into lower level representation

- EUROPHRAS: WS on multi-word units in MT and Translation Technologies, 2013on
- - by European association for phraseology and ACL

Connection to my PhD thesis

- **Challenges of Multi-word Expressions in Neural MT context**
  - Its unpredicted forms/syntax: by and large (prep+conj+adj)
  - Its semantic/figurative interpretation: idioms, metaphors (kick the bucket)
  - Its broad category with (semi-/)fixed expressions: verbal MWEs, light verb constructions, compound words, named entities, etc.
- **The study of MWEs are expected to facilitate development of better NMT models**
  - Reduce the ambiguity - word/phrase sense disambiguation
  - Correctly capture the syntax and semantics of terms and context
  - Indicators on how MTs progress, and the ways they should

- "An investigation into multi-word expressions in machine translation" ADAPT centre/DCU, Ireland

The hypotheses of thesis

- **Hy: MWEs set challenges for MT and investigation into MWEs can help to improve MT from both translation modelling and translation evaluation perspectives.**

- Hy.a: Current methodologies addressing MWEs in MT can be improved to include language-specific characteristics and features.

- Hy.b: MWEs in MT can be addressed using technologies inspired by tackling rare word/phrase and OOV (out-of-vocabulary) words issues, from statistical point of view.

- Hy.c: MT quality assessment models and test suites need to be redesigned and recreated to be able to reflect the translation accuracy of idiomatic MWEs.

What were the proposed Research Questions (RQs)?

- RQ-I: We assume that the proper integration of MWE knowledge into MT models can help to improve translation quality of MWEs in the source text and further improve the overall text/context translation

  - a) Validate and re-examine the state-of-the-art of existing models of integrating MWEs knowledge into NMT but with different language pairs of high-resource and high-performance baseline setting (de/zh-en).

  - b) Design new methods for integrating MWE knowledge into MT by addressing weak points of existing models, focusing on exact language pairs (zh-en).

  - c) Verify the proposed models with quantitative and qualitative evaluations, having human experts' validation wherever possible.

What were the proposed Research Questions (RQs)?

- RQ-II: We assume a new test suite with MWE annotations can be created for better assessment of current state-of-the-art MT models, and an improved MT evaluation methodology can be achieved with MWEs taken into account (e.g. human-in-the-loop workflow).
  - a) Revisit the problems and challenges in MT modelling and MT assessment, perform a critique review.
  - b) Create a corresponding new test suite as multilingual resource with MWE annotations for better MT quality assessments/examinations towards real human parity.
  - c) Address the issues in conventional MT assessment methods, both human judgement and automatic metrics, and design new methodologies that incorporate MWEs into the evaluation system/framwork.

MT example: MWEs themselves and their influence of contextual words

ZH source:	年年歲歲花相似，歲歲年年人不同
ZH pinyin:	Nián nián suì suì huā xiāng sì, suì suì nián nián rén bù tóng.
EN reference:	The flowers are similar each year, while people are changing every year.
EN MT output:	One year spent similar, each year is different

2018 googleNMT

Chinese compound nouns as (time) adverbial (年年歲歲 nian nian sui sui, every year), word sense disambiguation (花 hua, flower/spend), NMT adequacy (人 ren, human/people)

Lifeng Han, Gareth J.F. Jones and Alan F. Smeaton. 2020. MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2970–2979 Marseille, 11–16 May 2020

MT example: MWEs themselves and their influence of contextual words

ZH source:	小明去學校上課了
ZH pinyin:	Xiǎo míng qù xué xiào shàng kè le
EN reference:	Xiao Ming went to school to attend classes
EN MT output:	Xiao Ming went to school

2018
googleNMT

Chinese compound nouns as (time) adverbial (年年歲歲 nian nian sui sui, every year), word sense disambiguation (花 hua, flower/spend), NMT adequacy (人 ren, human/people)

Lifeng Han, Gareth J.F. Jones and Alan F. Smeaton. 2020. MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2970–2979 Marseille, 11–16 May 2020

MT example: MWEs themselves and their influence of contextual words

ZH source:	燕雀安知鴻鵠之志哉？
ZH pinyin:	Yàn què ān zhī hóng hú zhī zhì zāi?
EN reference (literal):	How can a finch know the ambition of a big bird (or swan)?
EN MT output:	What is the meaning of Yanque Anzhihong?
EN reference:	Common folk do not know the ambitions of the very motivated people.

2018 googleNMT

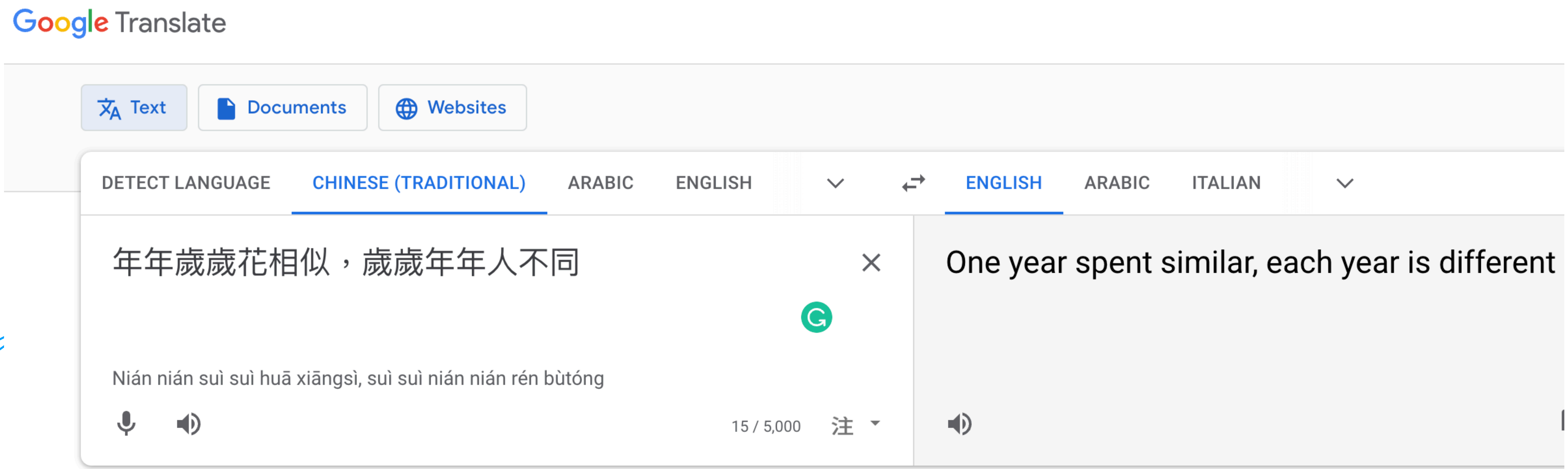
Chinese compound nouns as (time) adverbial (年年歲歲 nian nian sui sui, every year), word sense disambiguation (花 hua, flower/spend), NMT adequacy (人 ren, human/people)

Lifeng Han, Gareth J.F. Jones and Alan F. Smeaton. 2020. MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2970–2979 Marseille, 11–16 May 2020

After four years? 2018-2022

- What happened during these four years?
- Google multilingual NMT
- PLMs for NMT, Large/extra-large Multilingual PLMs for NMT, Meta-AI's team (mBERT50, M2M-100, NLLB-200)

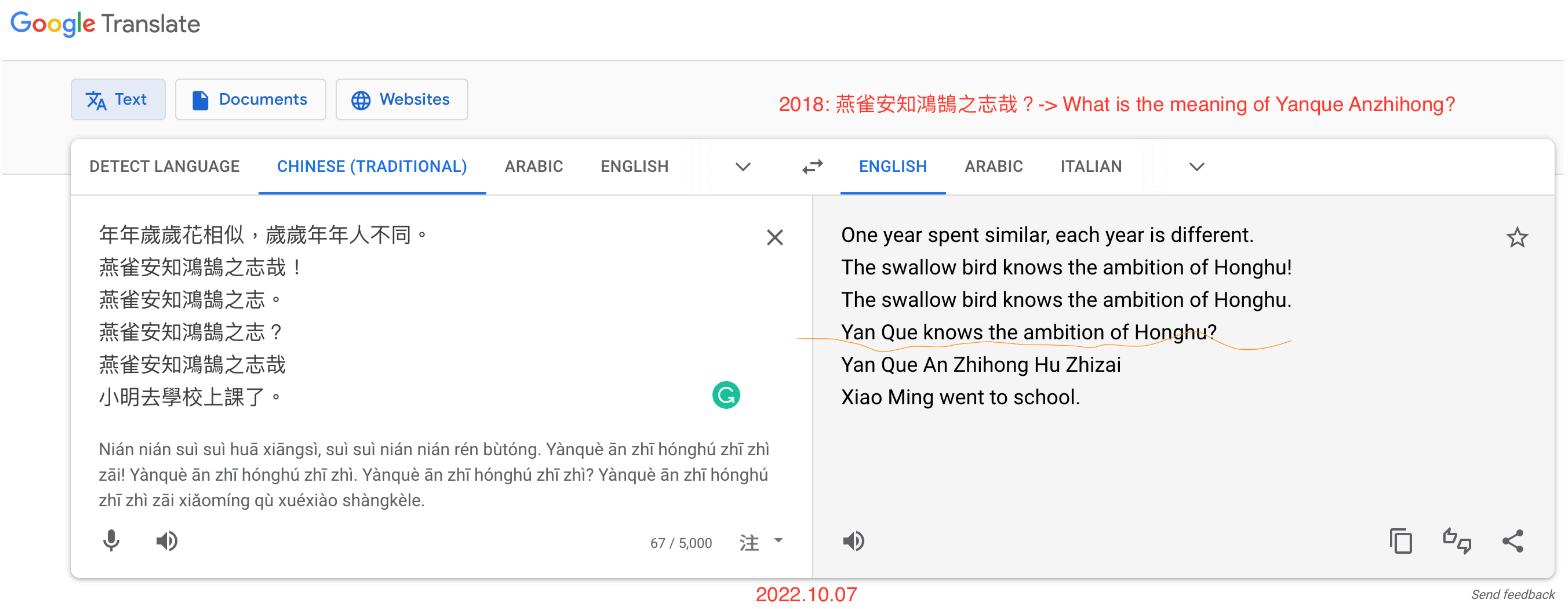
Can you believe?



2022.10
googleNMT

Exactly the same translation output.
Integrating huge amount of data to their model did not help: on MWEs? how to better use big data?
Domains/literature/news/social-media/biomedical/clinical/law? How about mixed domain/genres expressions/context MT? e.g. this sentence can appear in a modern Chinese context.
Multilingual models did not help: MWEs? though it might helped low resource languages; but for “high-resource” and challenge MT issues, it idid not work? Or we need to address low frequency-terms in high-resource languages?

Can you believe?



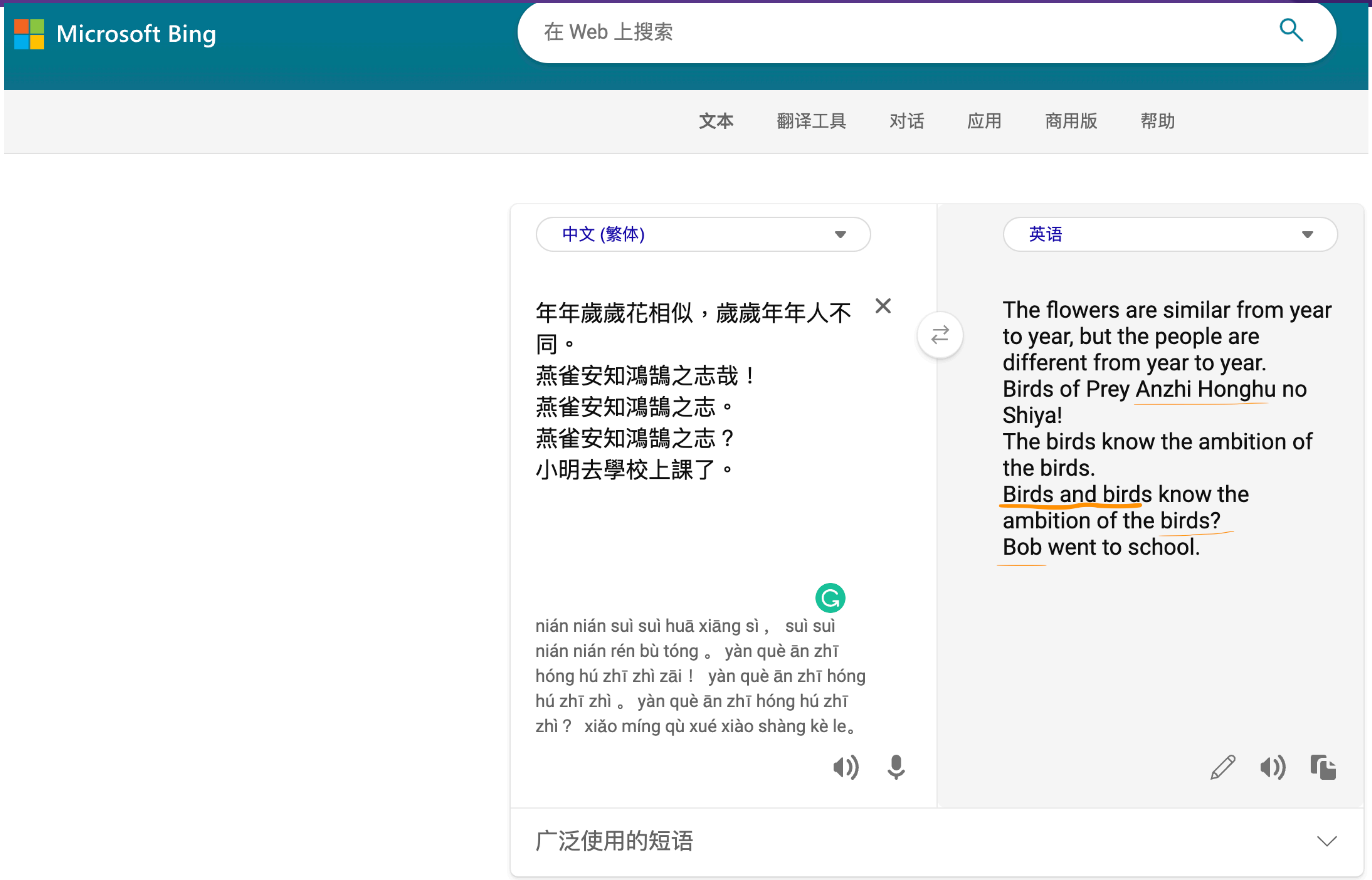
2022.10
googleNMT

Exactly the same translation output.

Integrating huge amount of data to their model did not help: on MWEs? how to better use big data? Domains/literature/news/social-media/biomedical/clinical/law? How about mixed domain/genres expressions/context MT? e.g. this sentence can appear in a modern Chinese context.

Multilingual models did not help: MWEs? though it might helped low resource languages; but for “high-resource” and challenge MT issues, it idid not work? Or we need to address low frequency-terms in high-resource languages?

Bing Translator?



www.adaptcentre.ie

Baidu Translator?

DeepL Translator?

DeepL 翻译器 DeepL Pro 为何选择DeepL? API 不同计划和价格 应用程序 免费 开始免费试用 登录

中文（自动检测） 英语（美式） 术语表

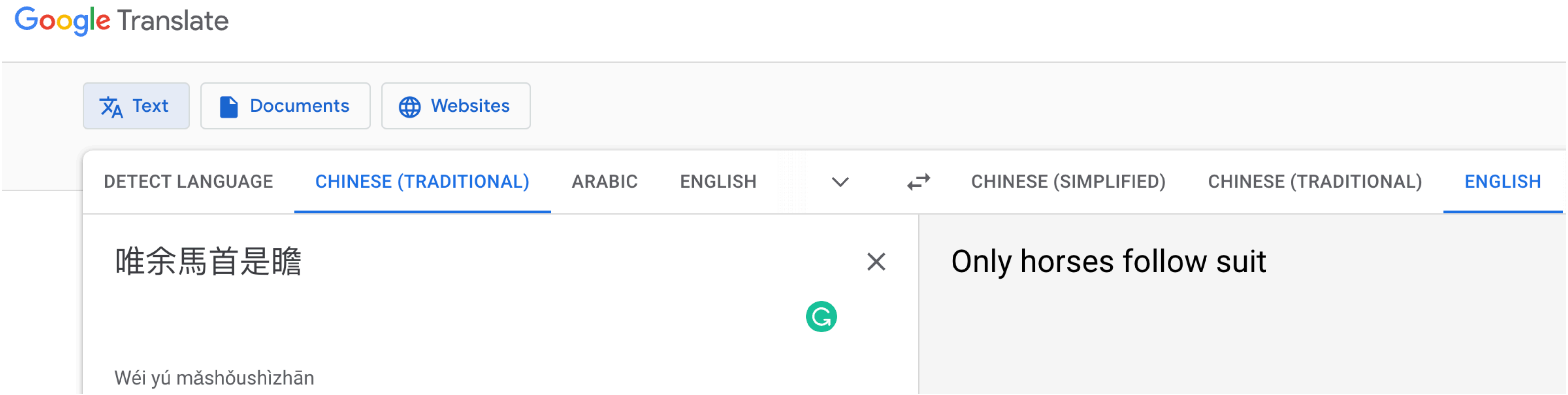
年年歲歲花相似，歲歲年年人不同。
燕雀安知鴻鵠之志哉!
燕雀安知鴻鵠之志。
燕雀安知鴻鵠之志?
燕雀安知鴻鵠之志哉
小明去學校上課了。

The flowers are similar every year, but the people are different every year.
How can a bird know the will of a great bird?
How can a bird know the will of a big bird?
How can a bird know the will of a big bird?
How can a bird know the will of a big bird?
The first thing you need to do is to go to school.

词典

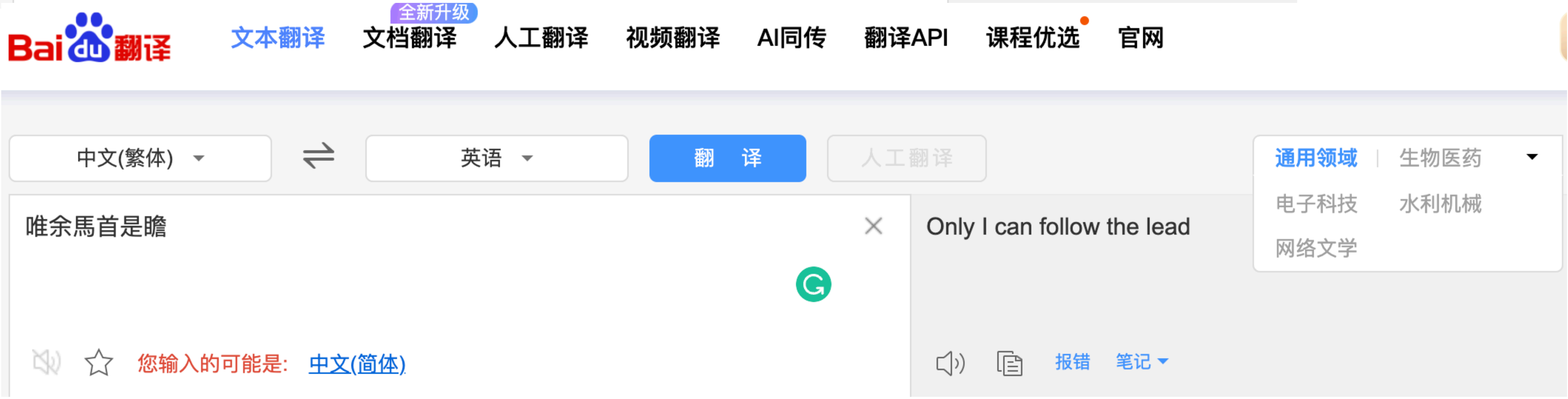
How did it make it one step closer than other NMTs? Even though the entities are not translated correctly?
How did it make the simpler sentence complicated and add extra noise/content?

How about a new example?



Bing Translator:

How about a new example?



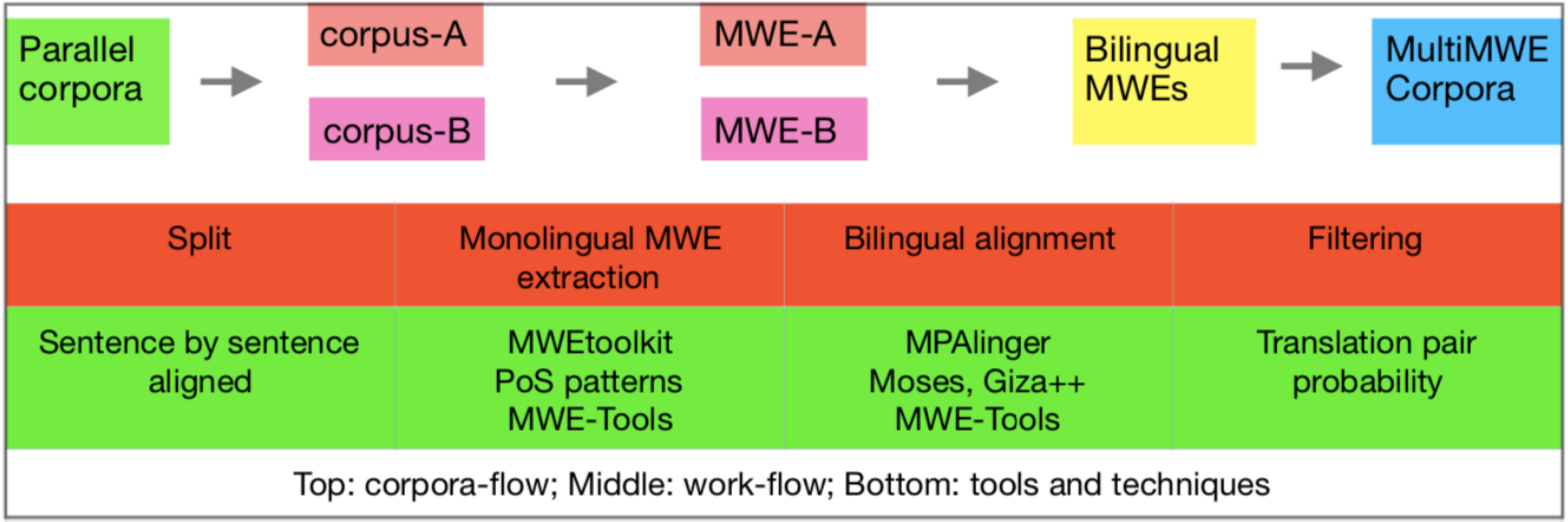
Look
at the domain
choice

唯...是瞻 (wéi … shì zhān, only look at …) - dis-continuous MWEs
唯<u>余马首</u>是瞻 (wéi yú mǎ shǒu shì zhān, only look at my horse’s head, only follow my lead.)

You thoughts on these examples?

BiMWE/data Argumentation

- Bilingual MWEs extractions and integration into NMT training



Lifeng Han, Gareth J.F. Jones and Alan F. Smeaton. 2020. MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2970–2979 Marseille, 11–16 May 2020

BLEU doesn't work really well

	n-gram scores				Params		Combine
models	1-gram	2-gram	3-gram	4-gram	BP	ratio	overall
Baseline	**63.3**	**35.2**	**21.4**	**13.5**	0.942	0.944	26.73
MWEpruned0.7+Base	63.0	35.1	21.3	**13.5**	**0.952**	**0.953**	**26.87**
ExterMWE871+Base	**63.3**	**35.2**	21.2	13.3	0.929	0.932	26.15

Table 4.2: DE-2-EN NMT BLEU Scores with 20k Transformer Learning Steps

	n-gram scores				Params		
models	1-gram	2-gram	3-gram	4-gram	BP	Ratio	overall
Baseline	**56.3**	**26.5**	14.3	8.2	**0.9**	**0.905**	18.39
MWE+Base	55.9	26.1	14.3	8.2	0.884	0.89	17.99
MWEpruned0.85+Base	55.9	26.3	**14.5**	**8.4**	0.899	0.903	**18.49**

Table 4.3: Zh-2-En NMT BLEU scores with 20k Transformer learning steps.

Human evaluations

	Examples of MWE translations in MT outputs
Src	俄罗斯 与 土耳其 领导人 周二 进行 会见 ， 双方 握手 并 宣布 正式 结束 长 达 八个 月 的 口水战 与 经济制裁 。
Ref	the leaders of Russia and Turkey met on Tuesday to shake hands and declare a formal end to an eight - month long war of words and economic sanctions .
Base	Russian and Turkish leaders met Tuesday , shaking hands and declaring the official end of eight months of water fighting and economic sanctions .
B+MWE	Russian and Turkish leaders met on Tuesday and both shook hands and announced a formal end of eight months of oral combat and economic sanctions .
Src	来自 所谓 朋友 的 攻击 更让人 难以 接受
Ref	the offence was even greater , coming from a supposed friend .
Base	attacks from a friend are even harder to accept .
B+MWE	the attack from so-called friends is harder to accept .

Human eval: looking into test set outputs
Even though BLEU scores do not reflect much score-gain.

MWE: 口水战 (kǒushuǐ zhàn, war of words)
- Base: “water fighting”
- B+MWE: “oral combat” (BLEU fails by simple n-gram)

MWE: 所谓 朋友 (suo wei peng you, supposed friend/s)
- Base: friend (adequacy loss)
- So-called friends (BLEU fails again, even favour Base)
- single reference issue and n-gram matching issue

* Alternative MT evaluation methods, or
* Better test suits with *multi-ref* or semantically paraphrased ref are required

Decomposing Chinese Symbols

劍 (jiàn) ↓

鋒 (fēng) ↓

Level-1: (phonetic, qiān) 僉刂 (semantic, *knife*) | (semantic, *metal*) 金夆 (phonetic, féng)

Level-2: 亼吅从 刂 | 人王丷 夂丰

Level-3: 人一口口人人 刂 | 人一土丷 夂三丨

…

… … | … …

Full-stroke: 丿㇏一丨㇕一丨㇕一丿㇏丿㇏ 丨亅 | 丿㇏一一丨丶㇀一 ㇀㇇㇏一一一丨

Word level	28 / 歲 / 廚師 / 被 / 發現 / 死 / 於 / 舊金山 / 一家 / 商場
Character	28 歲 廚 師 被 發 現 死 於 舊 金 山 一 家 商 場
Pronunciation	èr shí bā Suì chú shī bèi fā xiàn sǐ yú jiù jīn shān yī jiā shāng chǎng
Radical	28 止戌少 广尌 𠂤帀 衤皮 癶殳 王見 歹匕 方令 萑臼 人王丷 山 一 宀豖 亠丷冏 土昜
English Ref.	28-Year-Old Chef Found Dead at San Francisco Mall

Lifeng Han, et al. 2021. Chinese Character Decomposition for Neural MT with Multi-Word Expressions. Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), 336–344.

Combine word/character(symbol)/radical

	1-gram	2-gram	3-gram	4-gram
Baseline	0.6451	0.4732	0.3508	0.2630
W+C+R	**0.6609**	**0.4839**	**0.3572**	**0.2655**
W+C	0.6391	0.4663	0.3412	0.2527
W+R	0.6474	0.4736	0.3503	0.2607
C+R	0.6378	0.4573	0.3296	0.2410

Table 4.8: BLEU Scores on NIST08 Test Data

	1-gram	2-gram	3-gram	4-gram	5-gram
Baseline	5.1288	6.6648	7.0387	7.1149	7.1387
W+C+R	**5.2858**	**6.8689**	**7.2520**	**7.3308**	**7.3535**
W+C	5.0850	6.5977	6.9552	7.0250	7.0467
W+R	5.1122	6.6509	7.0289	7.1062	7.1291
C+R	5.0140	6.4731	6.8187	6.8873	6.9063

Table 4.9: NIST Scores on NIST08 Test Data

	Metrics Evaluated on 4-references		
Models	hLEPOR	BEER	CharacTER
Baseline	0.5519	0.4748	**0.9846**
W+C+R	**0.5530**	**0.4778**	1.3514
W+C	0.5444	0.4712	1.1416
W+R	0.5458	0.4717	0.9882
C+R	0.5353	0.4634	1.1888

Table 4.10: Broader Metrics Scores on NIST08 Test Data

* Evaluations using broader evaluation metrics on NIST MT challenge data.
*Radicals can enhance NMT learning for zh-en, as semantic features.

Replacing word sequence with decompositions

BLEU scores with increasing learning steps

	20k	100k	120k	140k	160k	180k
baseline	18.39	**21.56**	21.45	21.31	21.29	21.42
base+MWE	**18.49**	21.39	**21.67**	**21.83**	**21.42**	**21.86**
RXD3	16.48	20.75	20.73	20.93	20.98	21.14
RXD3+MWE	17.82	21.36	21.50	21.31	**21.42**	21.47
RXD2	11.84	13.26	12.88	13.02	13.38	12.86
RXD1/ideograph	15.52	20.67	20.61	21.26	20.76	21.00

src	28 岁 厨师 被 发现 死 于 **旧金山** 一家 **商场** 近日 刚 搬 至 旧金山 的 一位 28 岁 厨师 本周 被 发现 死 于 当地 一家 **商场** 的 **楼梯间** 。
ref	28 @-@ Year @-@ Old Chef Found Dead at **San Francisco Mall** a 28 @-@ year @-@ old chef who had **recently** moved to San Francisco was found dead in the **stairwell** of a local **mall** this week .
rxd3	the 28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef who **recently** moved to San Francisco has been found dead on a **stairwell** in a local **mall** this week .
base	the 28 @-@ year @-@ old chef was found dead in a **shop** in **San Francisco** a 28 @-@ year @-@ old chef who has moved to San Francisco this week was found dead on the **stairs** of a local **mall** .
base MWE	28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef who **recently** moved to San Francisco was found dead this week at a local mall .
rxd3 MWE	28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef **recently** moved to San Francisco was found dead this week at a local mall .
rxd1	the 28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef **recently** moved to San Francisco was found dead in a local shopping mall this week .
rxd2	the 28 @-@ year @-@ old chef was found dead in a **San Francisco mall** a 28 @-@ year @-@ old San Francisco chef was found dead in a local mall this week .

src	1. 旧金山 警察局 称该 起 死亡 案件 被 裁定 为 他杀 ， 并 正在 进行 调查 。 2. 该 发言人 称 ，“ 我们 是 相处 得 像 亲密无间 的 一家人 的 一个 小 团队 ， 我们 将 深深 怀念 他 。 ” 3. Sons & Daughters 餐馆 的 一位 发言人 表示 ， 他们 对于 Frank 的 死 感到 “ 非常 震惊 ” 。 4. 但 受害人 的 哥哥 表示 想不出 有 谁 会 想要 加害于 他 ， 并称 “ 一切 终于 好 起来 了 。 ”
ref	1. the San Francisco Police Department said the death was ruled a homicide and an investigation is ongoing . 2. " we are a small team that operates like a close knit family and he will be dearly missed , " the spokesperson said . 3. a spokesperson for Sons & Daughters said they were " shocked and devastated " by his death . 3.correction: a spokesperson for Sons & Daughters restaurant said they were " shocked and devastated " by Frank’s death . 4. but the victim 's brother says he can 't think of anyone who would want to hurt him , saying , " Things were finally going well for him . "
Base	1. the San Francisco police station says he was found dead and is investigating the death case . 2. " we are a small team that looks like a close @-@ knit family and we will miss him deeply , " the spokesman said . 3. a spokeswoman for Sons & Daughers said they were " shocked " by Frank 's death . 4. but the victim 's brother said he could not think of anyone who would harm him and said , " Everything is all right . "
base MWE	San Francisco police said the death was found to be a death case for him , and is under investigation . " we are a small team that is **as intimate as a family** , and we will remember him deeply , " the spokesman said . a Sons & Daughers spokesman said they were " shocked " at Frank 's death . but the victim 's brother said he couldn 't think of anyone who would want to hurt him and said , " **It 's all right at last** . "

Rxd1	the San Francisco police station said the death was ruled by him and was investigating . " we are a small team of people who live like a close family , and we will miss him deeply , " the spokesman said . a spokesman at the Sons & Daughers **restaurant** said they were " shocked " by Frank 's death . but the victim 's brother said he did not know who would want to harm him and said , " It 's all right . "
Rxd2	the San Francisco police station is investigating the death toll from a triangle . " we are a small team of families like the one with the same relationship , and we will miss him deeply , " the spokesman said . a spokesman for the Conversation & amp ; Conversation **Restaurant** said they were " very shocked " by the death of Frank . but the **elder brother** said that who would want to have the one who would , " all **is well at last** . "
Rxd3	the San Francisco police station said the death case was ruled by him and is under investigation . " we are a small group of close @-@ knit families and we will miss him deeply , " the spokesman said . a spokesman for Sons & Dauders said they were " shocked " by Frank 's death . but the victim 's brother said he could not imagine anyone trying to hurt him and said , " All is well . "
rxd3 MWE	San Francisco police said the death was convicted of **killing him** and was conducting an investigation . the spokesman said , " We are **as close as a family** of a small team , we will miss him deeply . " a spokesman for the Sons & Daughers **restaurant** said they were " shocked " by Frank 's death . but the victim 's brother said he could not think of anyone who would want to harm him , and said " **everything is better at last** . "

Abstract

In this talk, I will present some work related to multi-word expressions (MWEs), including MWE identification and translation. The presentation covers

1) our participation on **MWE identification** back in MWE2017@EACL,

2) our neural machine translation (NMT) models addressing **MWE translation from low-frequency terms** perspective for EN-DE/ZH, which includes bilingual MWE terms extraction and augmentation to training data, as well as decomposing Chinese characters/symbols into lower level representations, and

3) our **multilingual parallel corpus** preparation using PARSEME English seed corpus that has verbalMWE annotations, into DE/ZH/PL and other ongoing languages, where we give many examples when NMT went wrong in the task of translating MWE related content and MWE terms, and we try to classify these errors into different types to facilitate further research.

In the end of the talk, you will realise that when MT has reached a new level of quality via NMT, MWEs become very apparent bottlenecks in front of MT researchers.

We also propose some possible solutions to address these issues.

Construction of Multilingual MWE corpus

Parseme English → target languages → post-edited zh/pl/de/es… → Annotated MWEs → Validated Targets

MT compare four systems	Post-Editing of machine output sentences	Annotate target MWEs	Second round recheck from different person
We selected DeepL as our final MT output engine	Ensure fluency and adequacy, also paying attention to MWEs	The source and target MWEs are bilingual well aligned and annotated	To reduce human introduced / ignored translation errors

Top: corpora-flow; Middle: work-flow; Bottom: tools and techniques

Lifeng Han | Gareth Jones | Alan Smeaton. 2020. [AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations.]() Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons. 44–57.

NMT Error catergories facing MWEs

- EN-ZH:
  - Common sense
  - Super sense
  - Abstract phrases
  - Idioms
  - Metaphors
  - Ambiguities
    - Context-unaware ambiguity
    - social/literature-unaware ambiguity
    - Coherence-unaware ambiguity

- EN-DE/PL

HumanEval: MT on MWEs

Common sense

Source	At the corner of 72nd Street and Madison Avenue, he <u>waved down</u> a cab.	
DeepL	在72街和麦迪逊大道的拐角处，他向一辆出租车<u>招手</u>。 Zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù, tā xiàng yī liàng chū zū chē <u>zhāo shǒu</u>.	Wave hand
Bing	在72街和麦迪逊大道的拐角处，他<u>挥手示意</u>一辆出租车。 zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù , tā <u>huī shǒu shì yì</u> yí liàng chū zū chē.	Wave hand to indicate
Google	在第72街和麦迪逊大街的拐角处，他<u>挥舞着</u>出租车。 Zài dì 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā <u>huī wǔ zhe</u> chū zū chē.	Wave and dance with
Baidu	在72街和麦迪逊大街的拐角处，他<u>挥手叫了</u>一辆出租车。 zài 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā <u>huī shǒu jiào le</u> yí liàng chū zū chē.	Wave and got

HumanEval: MT on MWEs

Super sense

Source	Each time he took a walk, he felt as though he were leaving himself behind, and by giving himself up to the movement of the streets, by reducing himself to a seeing eye, he was able to escape the obligation to think, and this, more than anything else, brought him a measure of peace, a salutatory emptiness within.
DeepL	每一次散步，他都觉得自己好像把自己抛在了脑后，把自己交给了街道的运动，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这比什么都重要，这给他带来了一定程度的平静，内心的空虚让人敬佩。(jiāo gěi; nèi xīn de kōng xū ràng rén jìng pèi)
Bing	每次散步时，他都觉得自己好像要把自己抛在脑后，通过向街道运动自在，通过把自己留眼，他能够逃避思考的义务，这比什么都重要，给他带来了一定程度的和平，一种有益的空虚。(zì zài; yǒu yì de kōng xū)
Google	每次散步时，他都会感觉自己好像在抛弃自己，投身于大街小巷，睁大了眼睛，这使他摆脱了思考的义务，而这， 最重要的是，他带给他一种和平的感觉，一种内在的称呼空虚。(tóu shēn yú; nèi zài de chēng hu kōng xū)
Baidu	每次他散步，他都觉得自己好像是在离开自己，把自己交给街道的流动，把自己变成一个有眼光的人，他就可以逃避思考的义务，而这比其他任何事情都能给他带来某种程度的平和，一种内在的致意的空虚。(nèi zài de zhì yì de kōng xū)
Ref.	每一次散步，他都觉得自己好像把自己抛在了脑后，投身于大街小巷，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这给他带来了某种程度的平静和内心悦纳的空无，远胜于其他。(nèi xīn yuè nà de kōng wú)

Gave to

Self-enjoy

devoted himself to

An inner calling emptiness

Gave to

devoted himself to

An inner emptiness you are happy with

HumanEval: MT on MWEs

Abstract phrase - needing of context

Source	Quinn had his doubts, but this was all he had to **go on**, his only bridge to the present.
DeepL	奎恩有他的疑虑，但这是他唯一的依靠，是他通往现在的唯一桥梁。(wéi yī de yī kào) Only relying thing
Bing	奎因有他的怀疑，但这就是他必须**去**的，他唯一的桥梁到现在。 (bì xū qù de) Has to go
Google	奎因有他的疑惑，但这就是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de) Need to do
Baidu	奎因有他的疑虑，但这是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de) Need to do
Ref.	奎恩曾有他的疑虑，但这是他**开展工作**的所有依据，是他通往现在的唯一桥梁。(**kāi zhǎn gōng zuò** de suǒ yǒu yī jù) All he had to carry on the work / to work from (finding someone)

Sentence level MT would make no much sense in this situation.
- context aware MT, and document-level MT

Idiom - examples

Source	I was smoking my pipe quietly by my dismantled steamer, and saw them all cutting capers in the light, with their arms lifted high, when the stout man with mustaches came tearing down to the river, a tin pail in his hand, assured me that everybody was 'behaving splendidly, splendidly, dipped about a quart of water and tore back again.
DeepL	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下高举双臂切着荠菜，这时，那个留着胡子的壮汉撕扯着来到河边，手里拿着一个铁皮桶，向我保证说，大家都'表现得很精彩，很精彩，浸了大约一夸脱的水，又撕扯回来了。 (qiē zhe jì cài; sī chě huí lái) Cutting 'Shepherd's Purse'
Bing	Cutting shawl 我悄悄地用我拆除的蒸笼抽我的烟斗，看到他们所有的切割披肩在光中，他们的手臂抬起来高，当粗壮的人用胡子来到河中，一个锡桶在他的手中，向我保证，每个人都'表现辉煌，辉煌，浸约夸脱的水，并再次撕毁。(qiē gē pī jiān; zài cì sī huǐ)
Google	我正用拆卸的蒸锅安静地抽着烟斗，看到他们都在灯光下割下刺山柑，双臂高高举起，当时那留着小胡子的矮壮男人向河边撕下，手里拿着一个铁桶，向我保证 每个人都表现得非常出色，表现得非常出色，蘸了一夸脱的水，然后又撕了回来。(gē xià cì shān gān; sī le huí lái) Cutting 'capers'
Baidu	我正静静地在我拆掉的汽船旁抽着烟斗，看见他们都高举着胳膊，在灯光下割着山柑，这时那个长着胡子的胖男人手里拿着一个锡桶，朝河里跑来，向我保证每个人都表现得很好，很漂亮，蘸了一夸脱水，然后又往回跑。(gē zhe shān gān; wǎng huí pǎo) Cutting 'capers'
Ref.	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下欢呼雀跃，高举双臂，这时，那个留着胡子的大块头，手里拿着一个铁皮桶，快速来到河边，向我确保大家都"表现得很精彩，很精彩"，他浸了大约一夸脱的水，又快速回去了。(huān hū què yuè; kuài sù huí qù)

cut a caper: make a playful skipping movement

HumanEval: MT on MWEs

Metaphor examples

Source	The what? Auster laughed, and in that laugh everything was suddenly blown to bits. The chair was comfortable, and the beer had gone slightly to his head.
DeepL	那个什么？奥斯特笑了，在这笑声中，一切突然被炸得粉碎。(bèi zhà dé fěn suì) 椅子很舒服，啤酒已经微微到了他的头上。 (wēi wēi dào le tā de tóu shàng)
Bing	什么？奥斯特笑了，在笑，一切都突然被吹成位。(bèi chuī chéng wèi) 椅子很舒服，啤酒稍微到他的头去了。 (shāo wēi dào tā de tóu qù le)
Google	什么啊 Auster笑了起来，在那笑声中，一切突然被炸碎了。 (bèi zhà suì le) 椅子很舒服，啤酒微微飘到他的头上。(wēi wēi piāo dào tā de tóu shàng)
Baidu	什么？奥斯特笑了，在那笑声中，一切都突然被炸成碎片。(bèi zhà chéng suì piàn) 椅子很舒服，啤酒已经稍稍流到他的头上了。(shāo shāo liú dào tā de tóu shàng le)
Ref.	那个什么？奥斯特笑了，在这笑声中，一切突然化为乌有。(huà wéi wū yǒu) 椅子很舒服，啤酒已经微微让他上了头。 (wēi wēi ràng tā shàng le tóu)

Category-VI: Context-Unaware Ambiguity (CUA)

In this case, the context, i.e. the background information, is needed for correct translation of the sentence. For instance, see Figure 5.9. DeepL gives the translation "it did not give me time though", while Bing and GoogleMT give the same translation "it/this did not give me one day's time" and Baidu outputs a grammatically incorrect sentence. From the pre-context, we understand that it means the speaker "did not feel that is special to him" or "did not have affection of that" after *all the Mormon missionary's effort towards him*. Interestingly, there is a popular Chinese idiom (slang) that matches this meaning very well "不是我的菜 (bù shì wǒ de cài, literally *not my dish*)". From this point of view, the context based MT model deserves some more attention, instead of only focusing on the sentence level. When we tried to put all background context information as shown in Figure 5.9 into the four MT models, they produce as the same output for this studied sentence, as for sentence level MT. This indicates that current MT models still focus on sentence-by-sentence translation when meeting paragraphs, instead of using context inference.

HumanEval: MT on MWEs

Han, Smeaton, Jones (2021) AlphaMWE
https://aclanthology.org/2020.mwe-1.6

Context-Unaware Ambiguity (CUA)

Source	`But it did not give me the time of day.`
DeepL	但它并没有给我时间。(gěi wǒ shí jiān)
Bing	但它没有给我一天的时间。(gěi wǒ yī tiān de shí jiān)
Google	但这没有给我一天的时间。(gěi wǒ yī tiān de shí jiān)
Baidu	但它没有给我一天中的时间。(gěi wǒ yī tiān zhōng de shí jiān)
Ref.	但我没有感到这个对于我特殊 / 但这不是我的菜。(gǎn dào zhè ge duì yú wǒ tè shū / … wǒ de cài)
Context	An old Mormon missionary in Nauvoo once gripped my knee hard as we sat side by side, and he put his arm about me and called me "Brother." We'd only met ten minutes before. He took me to his good bosom. His eyes began to mist. I was a prospect, an exotic prospect in old tennis shoes and a sweatshirt. His heart opened to me. It opened like a cuckoo clock. But it did not …

Social/Literature-Unaware Ambiguity (SLUA), Domain Knowledge

Category-VI: Social/Literature-Unaware Ambiguity (SLUA)

In this case, social knowledge of current affairs from news, or literature knowledge about some newly invented entities / phrases is required in order to get correct translation output. For instance, Figure 5.10 includes two sentences, one from politics and another from literature.

In the first sentence, "de-gnoming" is a word from Harry Potter, invented

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

Social/Literature-Unaware Ambiguity (SLUA), Domain Knowledge

Source	The moment they know the de-gnoming's going on they storm up to have a look. Then someone says that it can't be long now before the Russians write Arafat off.
DeepL	他们一知道去核的事，就会冲上去看一看。 (qù hé) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**注销**。(**bǎ** ā lā fǎ tè **zhù xiāo**)
Bing	当他们知道去诺格明是怎么回事， 他们冲了起来看看。 (qù nuò gé míng) 然后有人说，现在俄罗斯人要不长了，就**把**阿拉法特**注销**了。(**bǎ** ā lā fǎ tè **zhù xiāo** le)
Google	当他们知道正在逐渐消失的那一刻，他们便冲上去看看。 (zhèng zài zhú jiàn xiāo shī) 然后有人说，不久之后俄罗斯人**将**阿拉法特**注销**。 (**jiāng** ā lā fǎ tè **zhù xiāo**)
Baidu	他们一知道德格诺明正在进行，就冲上去看一看。 (dé gé nuò míng) 然后有人说，俄国人很快就会把阿拉法特**一笔勾销**了。 (bǎ ā lā fǎ tè **yī bǐ gōu xiāo** le)
Ref.	一知道去地精的事在进行，他们就冲上去观看。 (qù dì jīng) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**下課 /** 让…下台。 (**bǎ** ā lā fǎ tè **xià kè;** ràng…xià tái)

Figure 5.10: MT issues with MWEs: social/literature-unaware ambiguity

Social/Literature-Unaware Ambiguity (SLUA), Domain Knowledge

by its author, to refer to the process of ridding a garden of gnomes, a small magical beast. Without this literary knowledge it is not possible to translate the sentence correctly. For instance, even though this sentence is from a very popular novel that has been translated into many languages, DeepL translated it as “去核 (qù hé, de-nuclear)”, Bing translated it as “去诺格明 (qù nuò gé míng, *de*-nuògémíng” where “nuògémíng” is a simulation of the pronunciation of “gnoming” in a Chinese way, Baidu translated it as “德格诺明 (dé gé nuò míng)” which is the simulation of the pronunciation of the overall term “de-gnoming”.

MT on MWEs

Category-VI: Coherence-unaware Ambiguity (CohUA)

Source	`Two months ago I had to have an operation for a serious complaint.`
DeepL	两个月前，我因为一次严重的**投诉**不得不做手术。(**tóu sù** … zuò shǒu shù)
Bing	两个月前，我不得不做一个严重的**投诉**手术。 (zuò … **tóu sù** shǒu shù)
Google	两个月前，我不得不接受一次手术以应对严重的**投诉**。 (jiē shòu yī cì shǒu shù … **tóu sù**)
Baidu	两个月前，我因为严重的**投诉**不得不动手术。 (**tóu sù** … dòng shǒu shù)
Ref.	两个月前，我因为一次严重的**症状**不得不做手术。(**zhèng zhuàng** … zuò shǒu shù)

Coherence of the sentence, natural language inference will help the disambiguation.

This kind of MWE ambiguity can be solved by the coherence of the sentence itself, for instance, the example in Figure 5.11. The four MT models all translated the vMWE itself “have an operation” correctly in meaning preservation by “做/接受/动手术 (zuò/jiē shòu/dòng shǒu shù)” just with different Chinese word choices. However, none of the MT models translated the “reason of the operation”, i.e., “complaint” correctly. The word complaint has two most commonly used meanings “a statement that something is unsatisfactory or unacceptable” or “an illness or medical condition” and all four models chose the first one. According to simple logic of social life, people do not need to “have an operation” due to “a statement”, instead their “medical condition” should have been chosen to translate the word “complaint”. Because of the in-

Human in the loop Eval Method: HiLMeMe

```
HiLMeMe: Algorithms

Input: src(i) as source sentence(i), tar(i) as candidate MT ouput for target sentence, ref(i) as reference translation of src(i), MWE_src(i.n) as MWEs in src(i), tran_MWE(i.m) as translations of MWE_src(i.n), MWE_ref(i.p) as MWEs in ref(i), alt_MWEs(i.q) as alternative correct target MWEs not included in MWE_ref(i.p), non_MWEs(i) as common words and correct translation but not using MWEs in tar(i), NULL indecating MWE_src(i.n) not transalted or lost in tar(i).

if tar(i) matches ref(i) in any degrees
    Point(i.I) = General(fluency, adequacy)
    Point(i.II) = MWE(α, β, γ, θ) where
        α++ if tran_MWE(i.m) = MWE_ref(i.p)
        β++ if tran_MWE(i.m) = alt_MWEs(i.q)
        γ++ if tran_MWE(i.m) = non_MWEs(i)
        θ++ if tran_MWE(i.m) = NULL
    Weight(i) = Φ(Sem,Gra,Idi,Amb,ϕ) where
        Sem++ if (MWE_src(i.n), MWE_ref(i.p)) meet semantics factor
        Gra++ if (MWE_src(i.n), MWE_ref(i.p)) meet grammar factor
        Idi++ if (MWE_src(i.n), MWE_ref(i.p)) meet idiomaticity factor
        Amb++ if (MWE_src(i.n), MWE_ref(i.p)) meet ambiguity factor
        ϕ = the value of MWE_src(i.n) weighting for src(i)

HiLMeME(i) = Point(i.I) + Weight(i) x Point(i.II)
```

Sentence-level
MWE-level

Using PsychoPy v3.0 platform www.psychopy.org

Initial platform for HiLMeMe (2be refined)

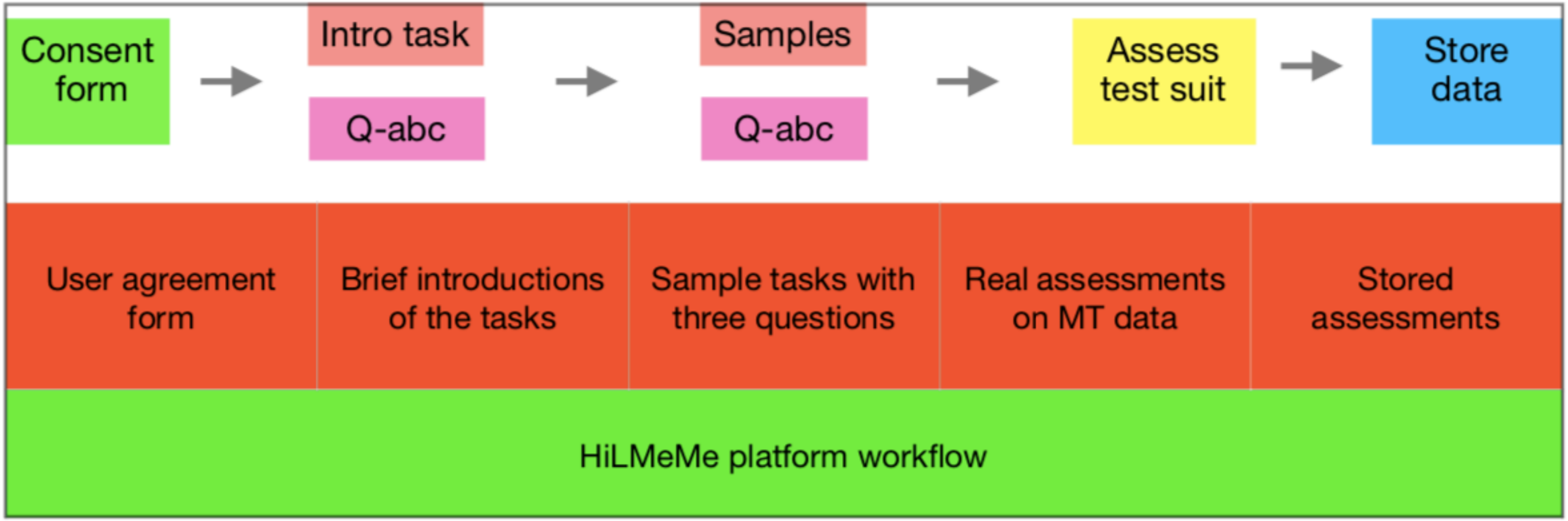


Figure 5.14: HiLMeMe Platform Workflow

MTE Platform HiLMeMe: **H**uman In the **L**oop **MT** **e**valuation looking into **M**ulti-word **E**xpressions

Looking back to the thesis

- Revisiting Hypotheses and RQs

- The **pilot** study and the **two-step character decomposition** experiments are an investigation into our hypotheses from the MWE for **MT** point of view, i.e., Hy.a and Hy.b, which followed the task list of RQ-I

  - => Verify the first half of hypotheses that addressing MWE issues can improve MT quality from MWE translation and the adequacy point of view, thus to improve overall MT performance.

  - => Inspiring the second half of hypotheses: new test-suits and evaluation methods are in demand.

- **MT errors analysis** and classification **facing MWEs** when creating AlphaMWE multilingual. & HiLMeMe **human in the loop MT evaluation** method looking into MWEs for Hy.c

  - => verify the importance of looking into MWEs in MT evaluation, the obstacles of MT meeting MWEs
  - => the needing of new test suits for MT assessment using reflecting idiomatic MWEs in translation
  - => reflecting the benefit: the human in the loop MT assessment workflow taking MWEs into account

How the work can be extended (viva time)

- Research directions in the future based this thesis

- RQ-I, Hy.a and Hy.b:
  - en-de/zh: Base+MWE model testing on semantic test suits: e.g. our own AlphaMWE corpus
  - Deeper decomposition of Chinese characters in MT, including full strokes order, on the performance of low-frequency words/phrases and MWEs.

- RQ-II, Hy.c:
  - Solid further experiments using HiLMeMe on MT evaluation
  - Human assessment agreements on this framework
  - Toolkit and platform refinement
  - Deploying HiLMeMe to broader application (NLP evaluation tasks)

- Other directions based on RQs & Hypotheses:
  - COVID19 domain medical data /health guideline MT covering new MWEs terms (our framework)
  - **Hybrid NMT model integrating both MWE compositionality detection and translation**

More Future work?

https://www.manchester.ac.uk/

- Hybrid NMT model integrating both MWE compositionality detection and translation
  - How a: constrained NMT with decoding part?
  - How b: extra-attention layer on MWEs for encoder?

- Can we use cross-lingual MWEs to measure similarity/classify languages/cultures?
  - The closeness of different languages via borrowing each other's MWEs? Culture influences?
  - The attempt to group languages into families from a new perspective? An evolution tree?

- Different kinds of MWEs => different solutions? MWE resources?

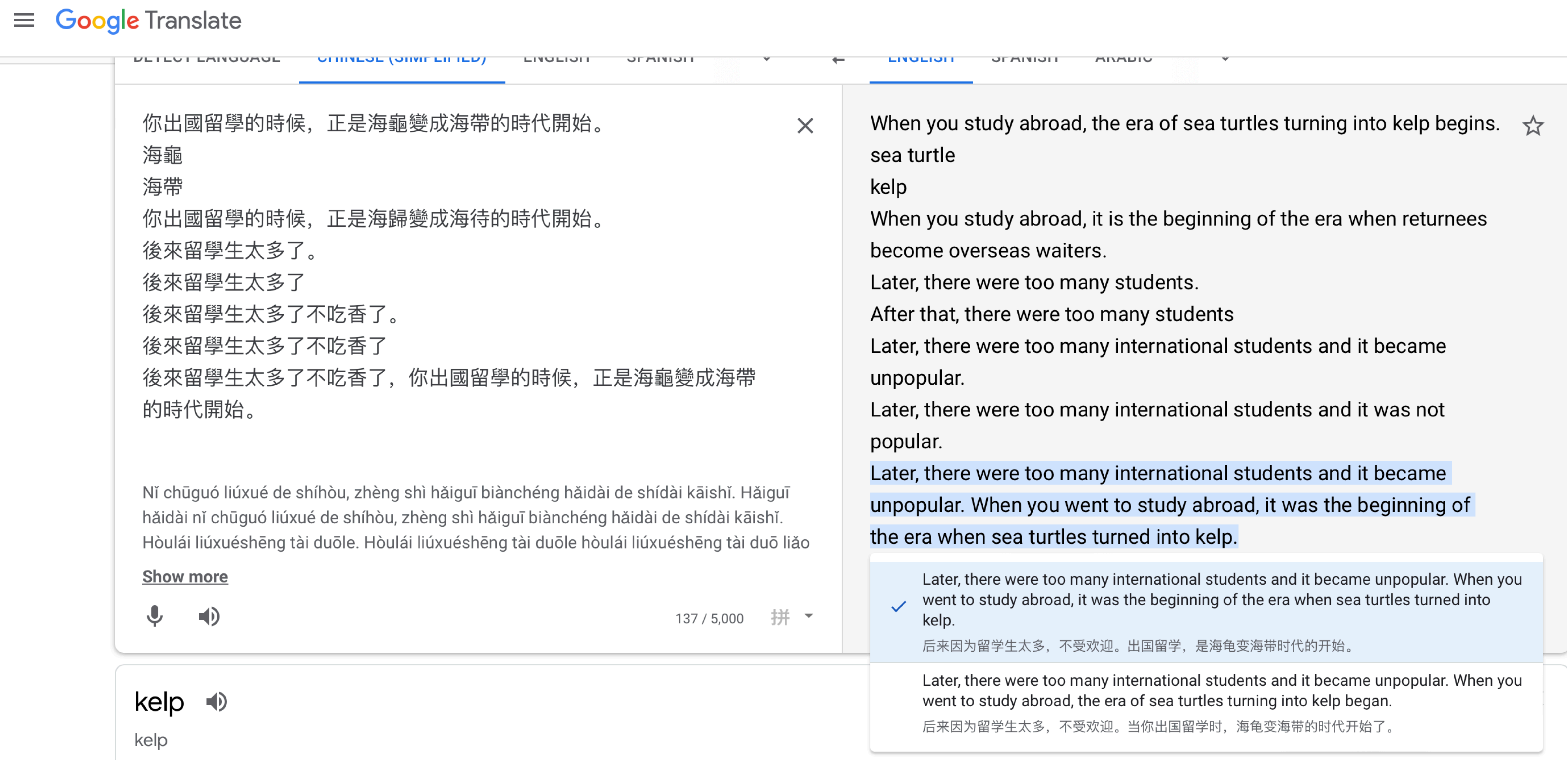
- How to take context into consideration when translating MWEs?
  - discourse/Paragraph/document-level NMT

MANCHESTER 1824

The University of Manchester

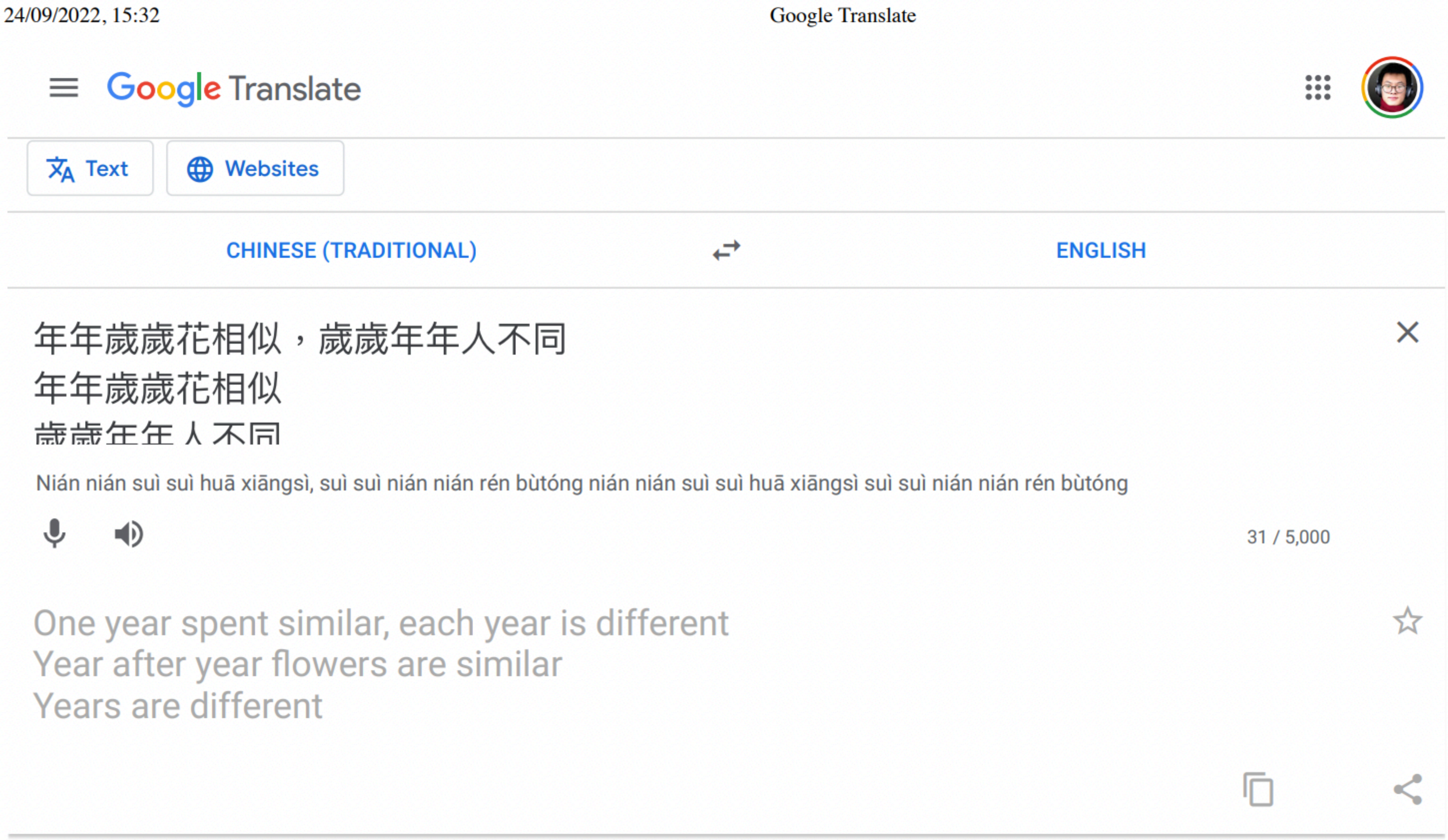
Appendices

Context does not matter? Why punctuation affects NMT?



On example: 年年歲歲花相似，歲歲年年人不同

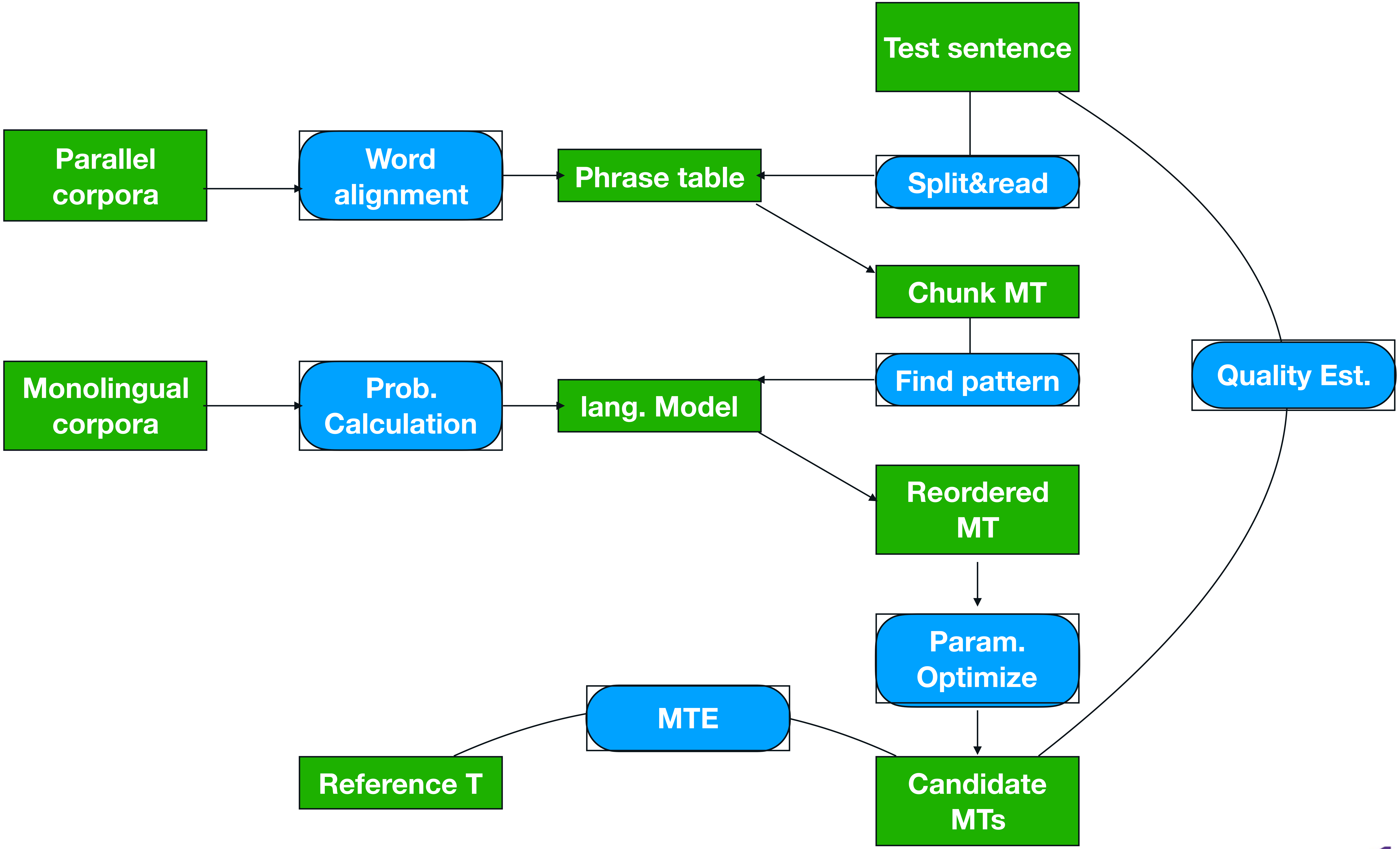
Breaking down the sentences, even more interesting



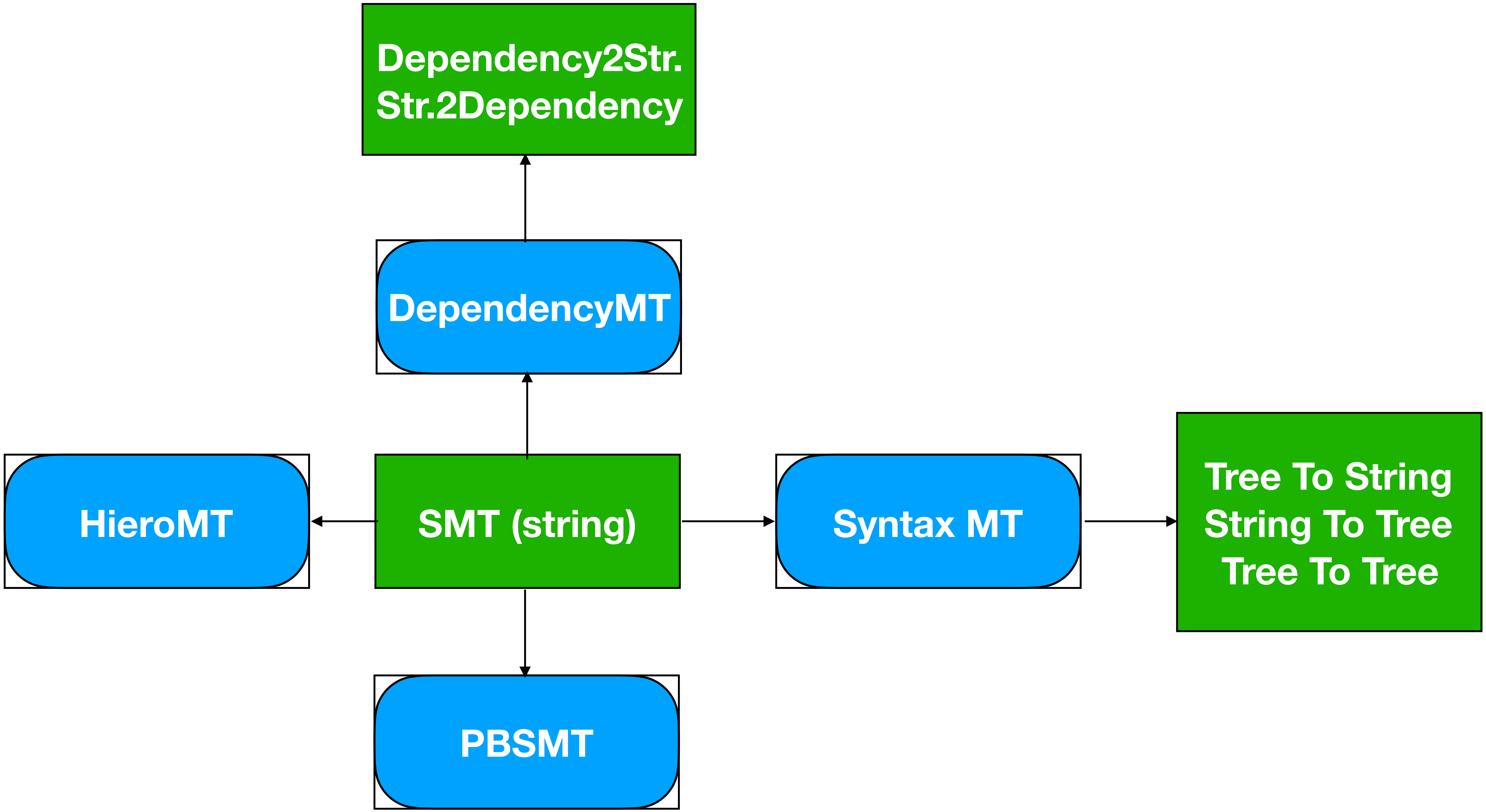
On MT structures

SMT vs NMT

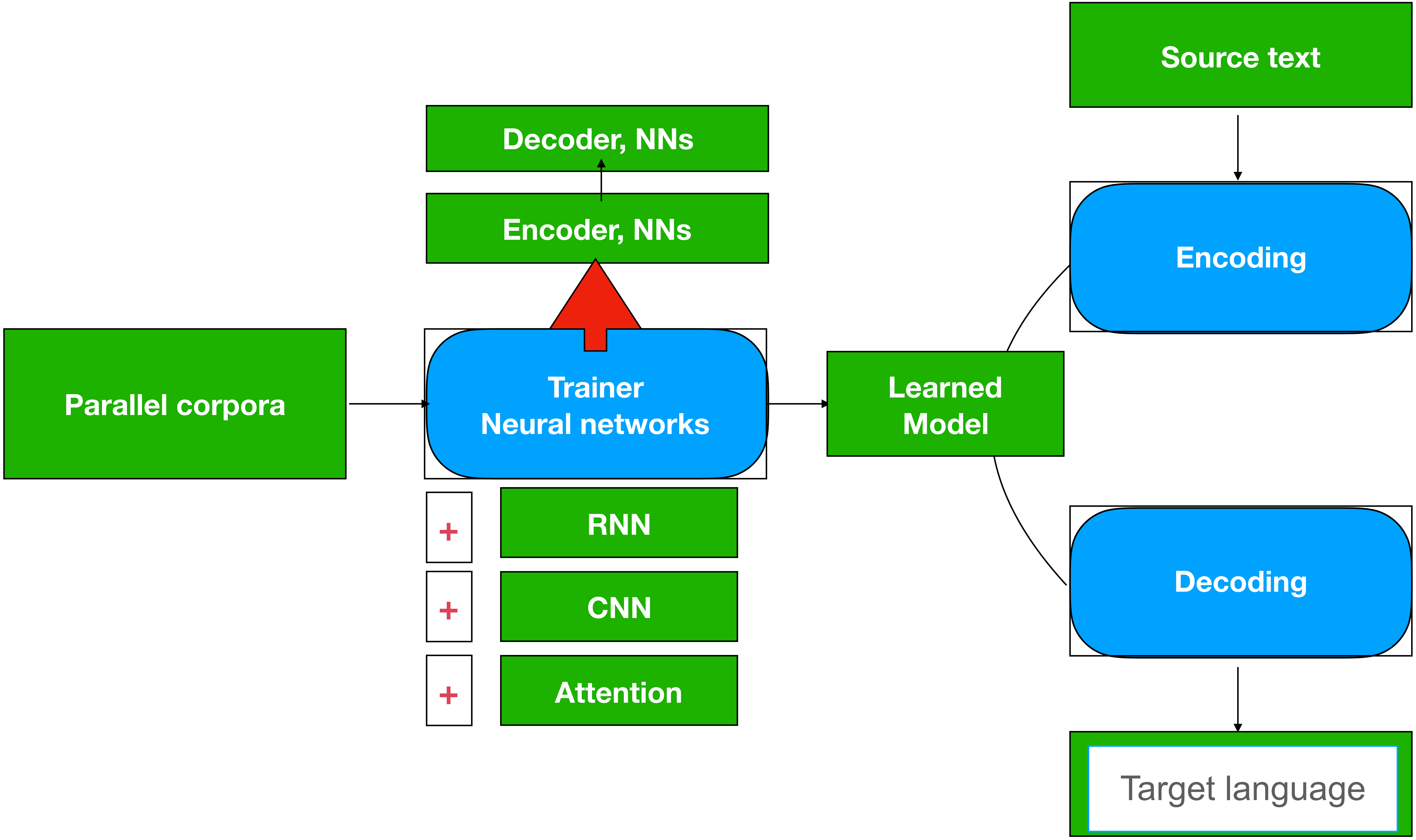
- Pros and Cons



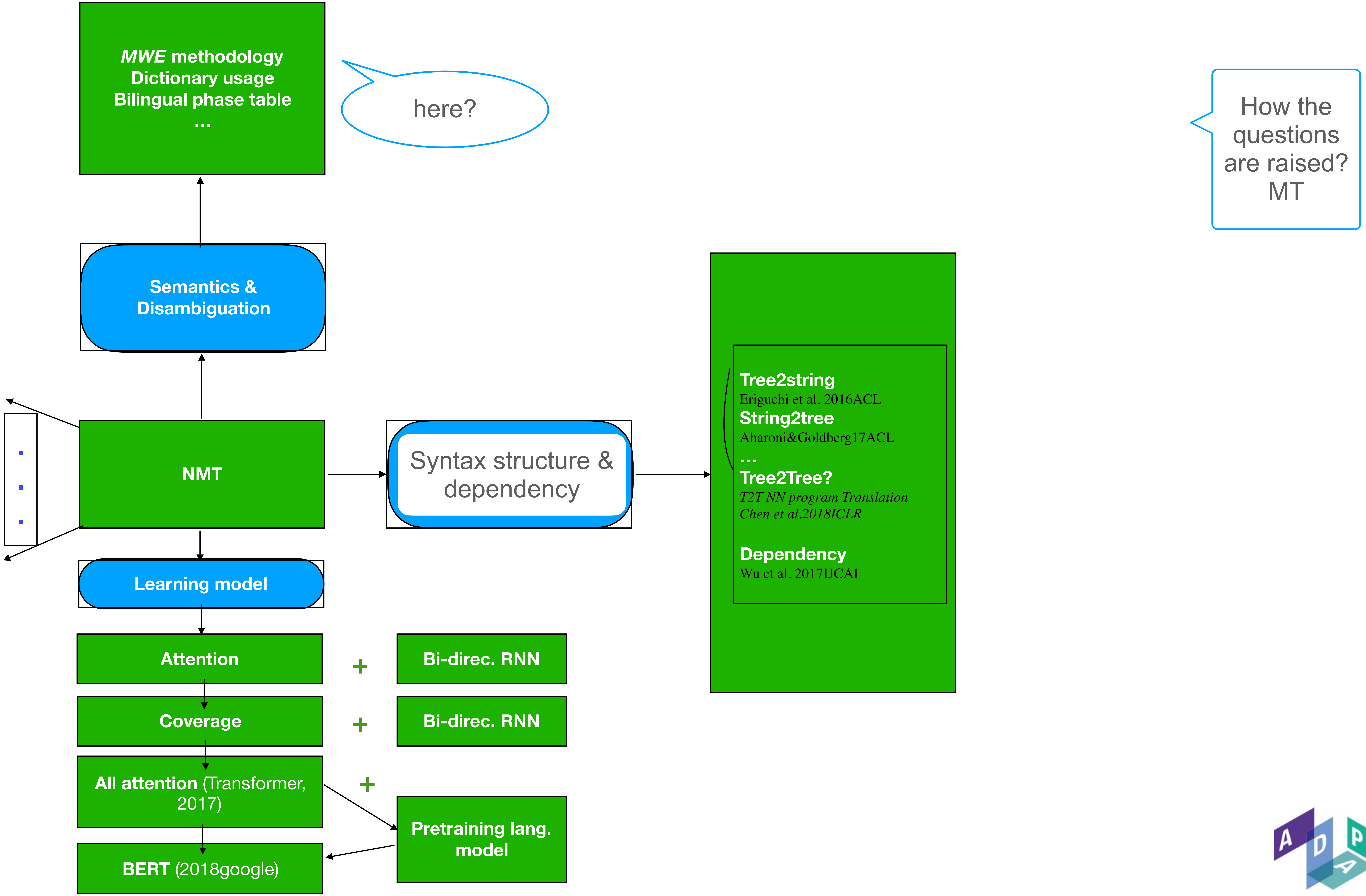
PBSMT components



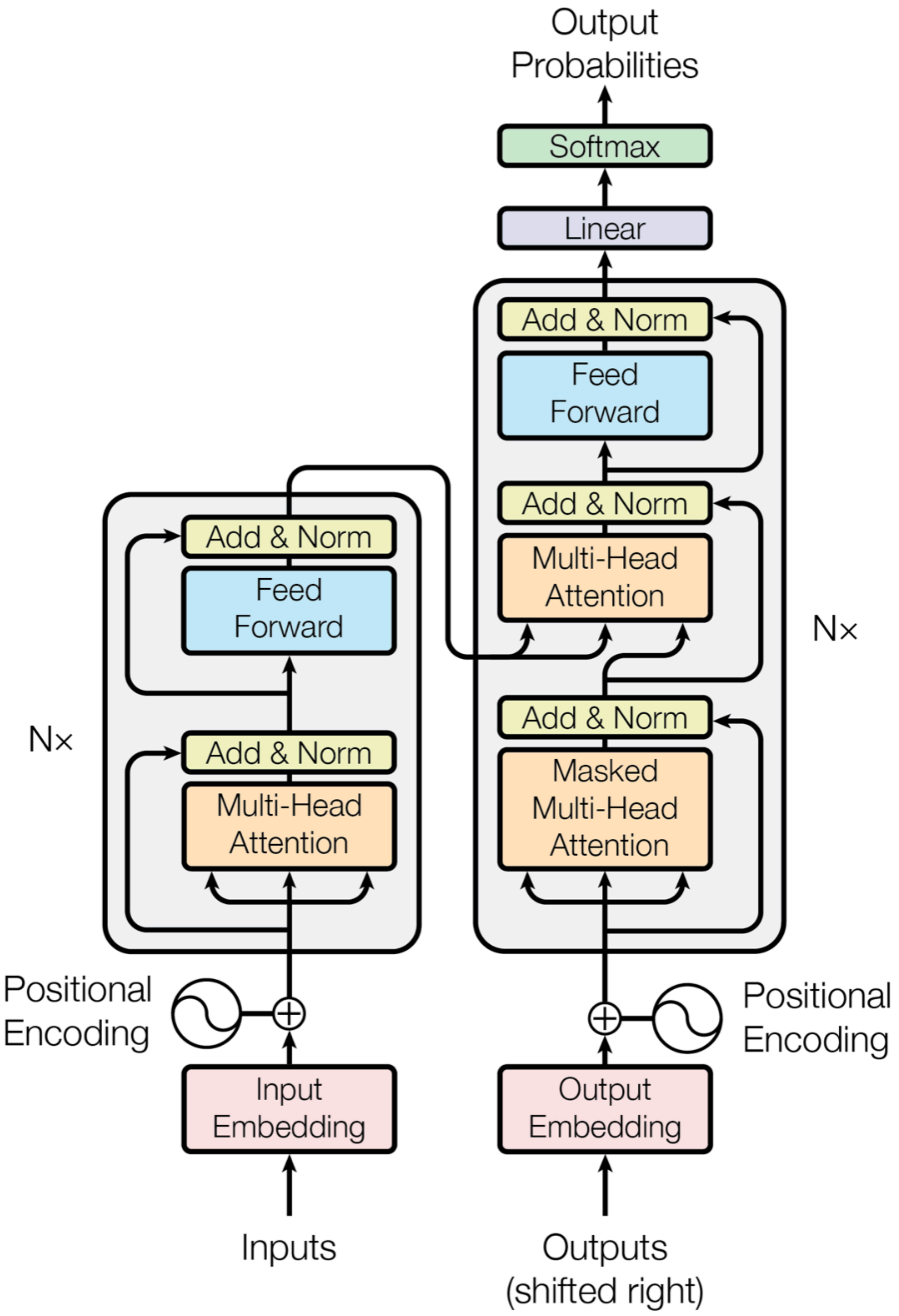
SMT branches



NMT components


MWE methodology
Dictionary usage
Bilingual phase table
...
here?
How the questions are raised? MT
Semantics & Disambiguation
NMT
Syntax structure & dependency
Tree2string
Eriguchi et al. 2016ACL
String2tree
Aharoni&Goldberg17ACL
...
Tree2Tree?
T2T NN program Translation
Chen et al.2018ICLR
Dependency
Wu et al. 2017IJCAI
Learning model
Attention
+
Bi-direc. RNN
Coverage
+
Bi-direc. RNN
All attention (Transformer, 2017)
+
Pretraining lang. model
BERT (2018google)

The Transformer structure
By Vaswani et al. (2017)
Main stream NMT



Related References

- Our earlier work related to this talk:
  - Han (2022) PhD **Thesis** 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559
  - NoDaLiDa21: "Chinese Character Decomposition for Neural MT with Multi-Word Expressions"

  - Lifeng Han, Gareth J. F. Jones, and Alan Smeaton. 2020. **MultiMWE**: Building a multi-lingual multiword expression (MWE) parallel corpora. In LREC.
  - Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. **AlphaMWE**: Construction of Multilingual Parallel Corpora with MWE Annotations. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 44–57, online. Association for Computational Linguistics.
  - MWE2017: "Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking" Maldonado, Han, Moreau, et al.
  - MWE Book chapter 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification

  - Gladkoff and Han. 2022LREC. **HOPE**: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation.
  - Measuring **Uncertainty** in Translation Quality Evaluation (TQE). 2022LREC. Gladkoff, Serge ; Sorokina, Irina ; Han, Lifeng ; Alekseeva, Alexandra
  - Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. **cushLEPOR**: customising hLEPOR metric using Optuna for higher agreement with human judgments or pre-trained language model LaBSE. In Proceedings of the Sixth Conference on Machine Translation, pages 1014–1023, Online. Association for Computational Linguistics.

  - Aaron Li-Feng Han and Derek F. Wong. (2016) Machine translation evaluation: A **survey** https://arxiv.org/abs/1605.04515 updated 2018 =>
  - Lifeng Han (2018) Machine Translation Evaluation Resources and Methods: A **Survey** https://arxiv.org/abs/1605.04515v8
  - Han et al. (2021) Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. https://aclanthology.org/2021.motra-1.3/
  - Lifeng Han (2022) An **Overview** on Machine Translation Evaluation. https://arxiv.org/abs/2202.11027 (in Chinese, English update forthcoming)
  - Lifeng Han. (2014) **LEPOR**: An Augmented Machine Translation Evaluation Metric. Msc **thesis**. University of Macau. https://library2.um.edu.mo/etheses/b33358400_ft.pdf
  - LREC22 tutorial "Meta-Evaluation of Translation Evaluation Methods: a systematic up-to-date overview" Han and Gladkoff https://github.com/poethan/LREC22_MetaEval_Tutorial

Codes and Platforms

- LEPOR: https://github.com/aaronlifenghan/aaron-project-lepor
- hLEPOR: https://github.com/poethan/LEPOR (older https://github.com/aaronlifenghan/aaron-project-hlepor)
- cushLEPOR: https://github.com/poethan/cushLEPOR
- HPPR: https://github.com/aaronlifenghan/aaron-project-hppr
- MultiMWE: https://github.com/poethan/MWE4MT
- AlphaMWE: https://github.com/poethan/AlphaMWE
- HOPE: https://github.com/lHan87/HOPE
- LREC22 tutorial: https://github.com/poethan/LREC22_MetaEval_Tutorial

Acknowledgement

- We thank our colleagues who are in contribution to our work and contributors who carried out corpus annotations and post-editing https://github.com/poethan/AlphaMWE .

- We thank the funding support from Uni of Manchester (via Prof Goran Nenadic's project) HIPS and JIGSAW, and previous funding from ADAPT Research Centre, DCU. Thanks also go to the IMS, University of Stuttgart for this funded trip as invited by Prisca. We thank the MWE17 slides prepared by Dr Alfredo Maldonado and detailed comments from Dr Erwan Moreau.