Paper Sharing

Zhu Liu

2023.05.10

Outline

- Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information
- A Psycholinguistic Analysis of BERT's Representations of Compounds

Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information

Niccolò Campolungo
Sapienza University of Rome
campolungo@di.uniroma1.it

Tommaso Pasini
Sapienza University of Rome
pasini@di.uniroma1.it

Denis Emelin
University of Edinburgh, Scotland
d.emelin@sms.ed.ac.uk

Roberto Navigli
Sapienza University of Rome
navigli@diag.uniroma1.it

NAACL 2022

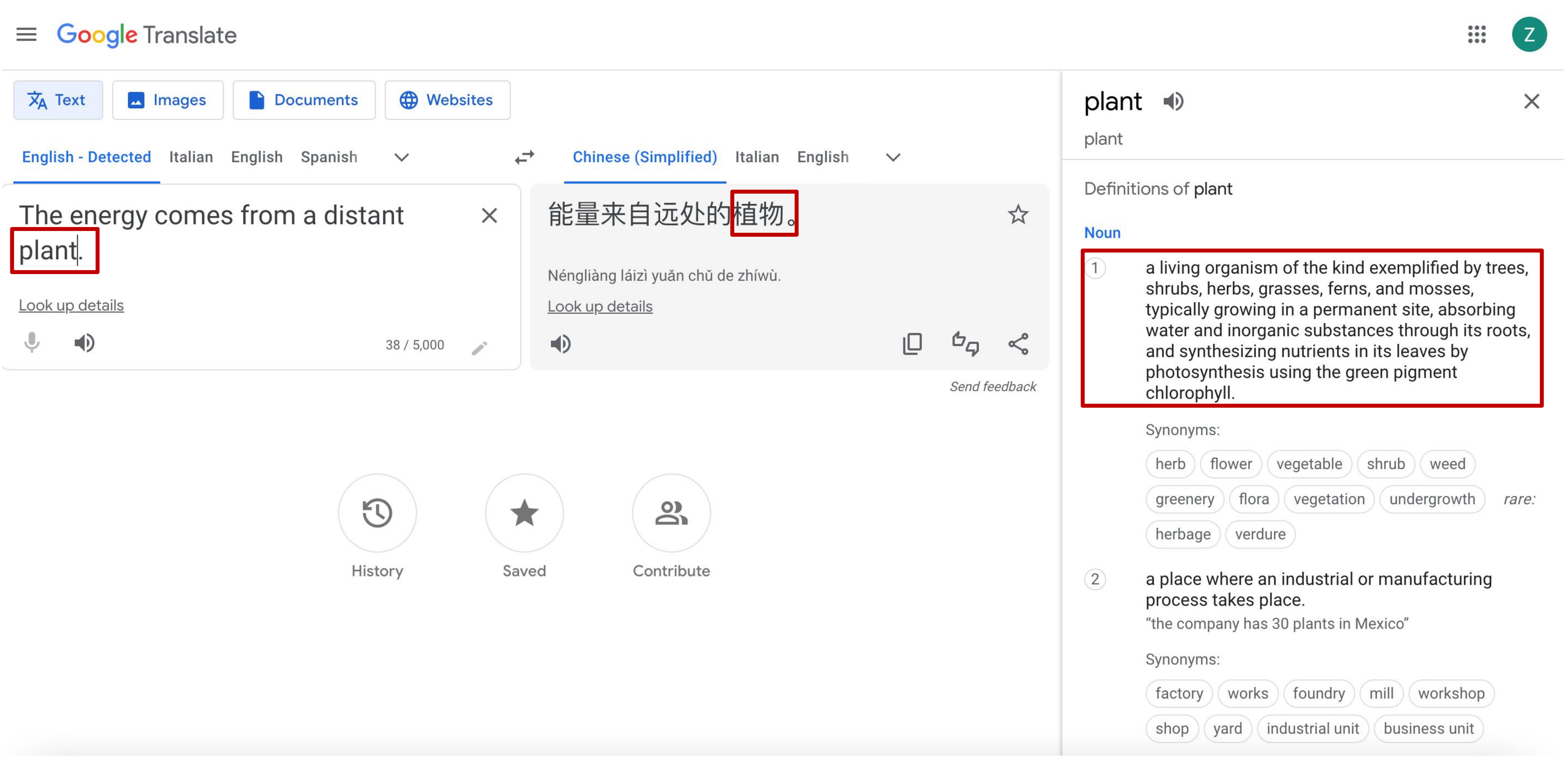
ACL 2023 & SapienzaNL

SAPIENZA
UNIVERSITÀ DI ROMA

- DMLM: Descriptive Masked Language Modeling
- What's the Meaning of Superhuman Performance in Today's NLU?
- Exploring Non-Verbal Predicates in Semantic Role Labeling: Challenges and Opportunitie
- RED^{FM}: a Filtered and Multilingual Relation Extraction Dataset
- Echoes from Alexandria: A Large Resource for Multilingual Book Summarization
- Incorporating Graph Information in Transformer-based AMR Parsing
- AMRs Assemble! Learning to Ensemble with Autoregressive Models for AMR Parsing
- Cross-lingual AMR Aligner: Paying Attention to Cross-Attention

Motivation

- Ambiguous words in Neural machine translation (NMT)
- MFS bias (most frequent sense bias) in NMT [Emelin et al., 2020]
- Inadequate measurement: Only BLEU
- NMT + explicit sense information, with challenges:
  1) scarce sense-tagged parallel data
  2) less accurate WSD systems until now
  3) Unclear how to incorporate them with neural models


Google Translate
Text
Images
Documents
Websites
English - Detected
Italian
English
Spanish
Chinese (Simplified)
Italian
English
The energy comes from a distant plant.
Look up details
38 / 5,000
能量来自远处的植物。
Néngliàng láizì yuǎn chù de zhíwù.
Look up details
Send feedback
History
Saved
Contribute
plant
plant
Definitions of plant
Noun
1
a living organism of the kind exemplified by trees, shrubs, herbs, grasses, ferns, and mosses, typically growing in a permanent site, absorbing water and inorganic substances through its roots, and synthesizing nutrients in its leaves by photosynthesis using the green pigment chlorophyll.
Synonyms:
herb
flower
vegetable
shrub
weed
greenery
flora
vegetation
undergrowth
rare:
herbage
verdure
2
a place where an industrial or manufacturing process takes place.
"the company has 30 plants in Mexico"
Synonyms:
factory
works
foundry
mill
workshop
shop
yard
industrial unit
business unit

Contributions

- Creating high-precision sense-annotated parallel corpora
- A fine-tuning strategy for incorporating these sense annotation
- Effective in both translation quality and bias disambiguation

Build a Sense-Annotated Parallel Corpus

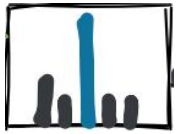
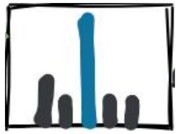
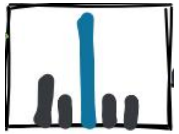
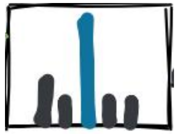
- Raw corpus: parallel sentences without word alignments and sense annotation

The energy comes from a distant plant

能源来自一个遥远的工厂

- Sense Scoring (annotation): WSD system
- Annotation Refinement: Word alignment
- BabelNet: A multi-lingual knowledge base

Build a Sense-Annotated Parallel Corpus

The energy comes from a distant plant

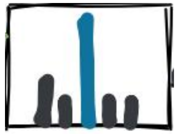
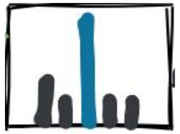
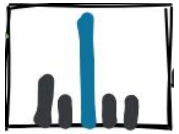
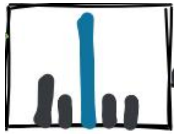
能源来自一个遥远的工厂

- Step 1: Sense Scoring (annotation):

A WSD system to classify each content word into the best BabelNet sense

$s = [w_1, \ldots, w_n]$ $\sigma(w_i)$ a score $c(S|w_i, s)$ to each synset $S \in \sigma(w_i)$

Build a Sense-Annotated Parallel Corpus

The energy comes from a distant plant
能源来自一个遥远的工厂

- Step 2: Annotation Refinement: Word alignment

  Aligned words have the same sense label

$$\mathcal{A} = \{(w_i^s, w_j^t) | w_i^s \in s, w_j^t \in t\}$$

$$P = (w_i^s, w_j^t) \in \mathcal{A}$$

$$\sigma(P) = \sigma(w_i^s) \cap \sigma(w_j^t)$$

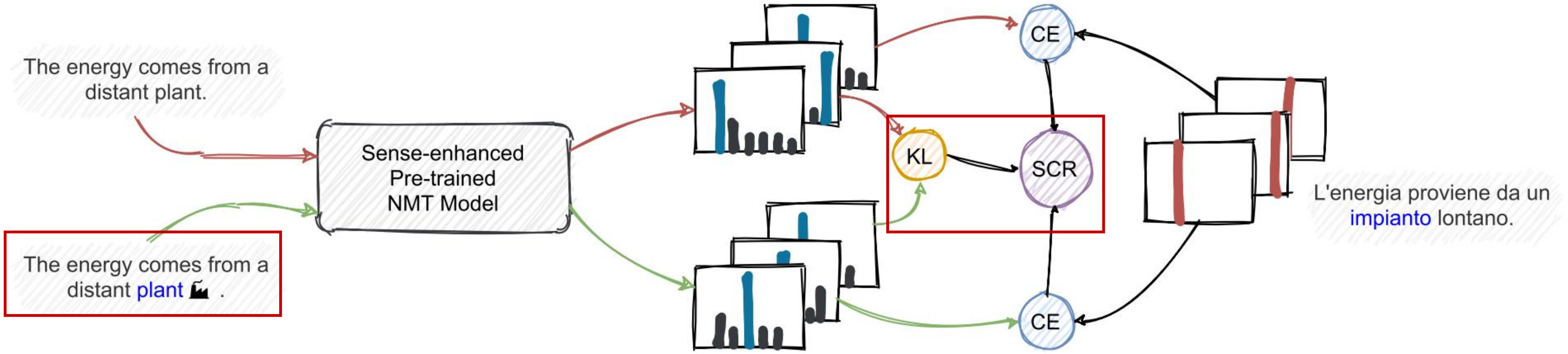
$$\sigma(P) = \emptyset \vee |\sigma(w_i^s)| < 2$$

$$S^* = S^*_{w_i^s} = S^*_{w_j^t} = \operatorname*{argmax}_{S \in \sigma(P)} \left(\frac{c(S|w_i^s, s)}{Z_s} + \frac{c(S|w_j^t, t)}{Z_t} \right)$$

$$Z_s = \sum_{S \in \sigma(P)} c(S|w_i^s, s)$$

$$Z_t = \sum_{S \in \sigma(P)} c(S|w_j^t, t)$$

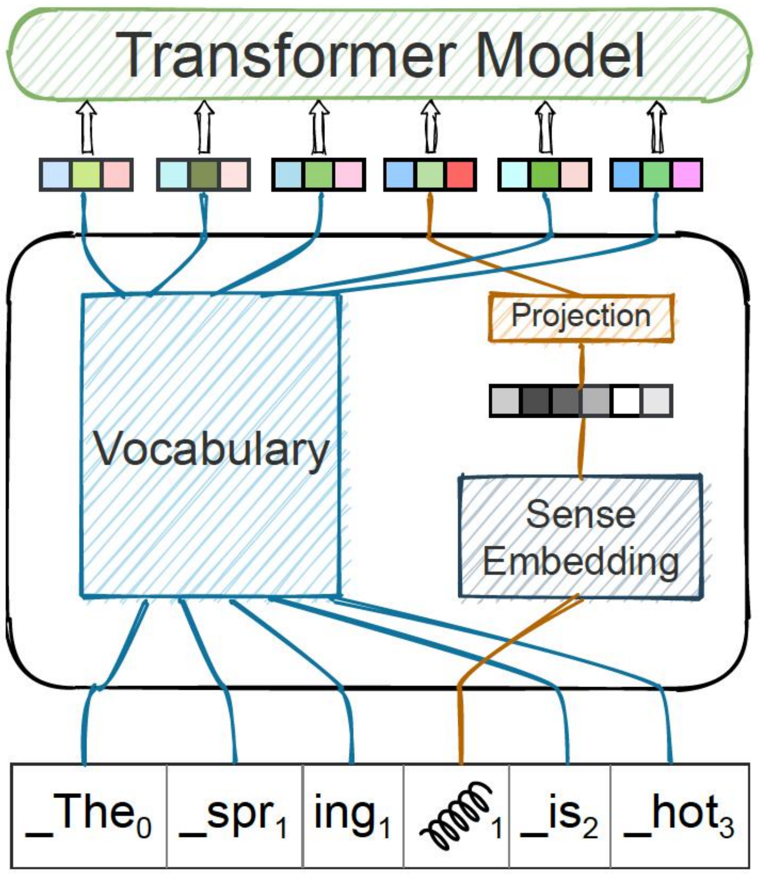
Semantic Injection



- Semantically Enhancing Sentences
- Semantic Consistency Regularization

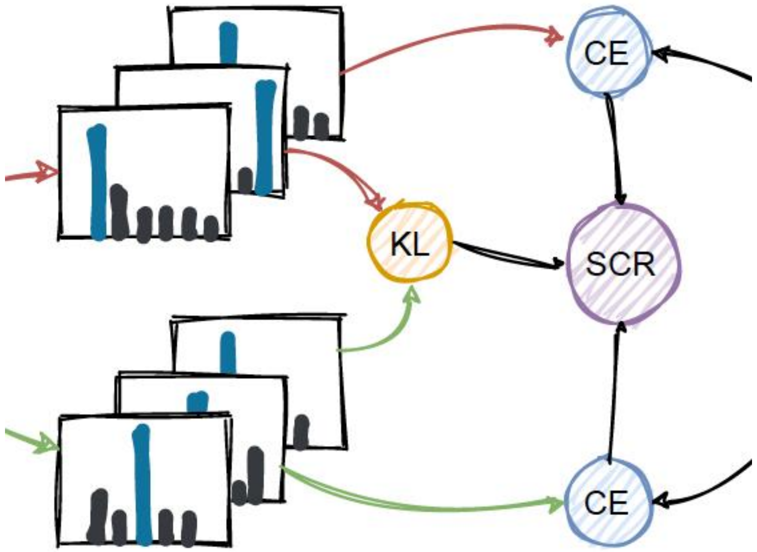
Semantically Enhancing Sentences

- Follow the target token,
- The same position id
- Sense embedding modular
- (Trainable) Projection Layer



Semantic Consistency Regularization

$$SCR(\theta) = -\log \mathcal{P}_\theta(y|x') - \log \mathcal{P}_\theta(y|x'') + \mathcal{D}_{\mathrm{KL}}(\mathcal{P}_\theta(y|x') \,||\, \mathcal{P}_\theta(y|x''))$$



- Teacher-student model (self-distillation process)
- Using Teacher (sense-aware data) only in training.
- ”partial/pseudo” supervised learning (The supervised signal is observable in training but hidden in inference)

Experiments

- Models

1) Translation model: 6 encoder layers + 6 encoder layers
2) (Pretrained) Sense embeddings: ARES [Scarlini et al., 2020b]
3) WSD model: EWISER

- Datasets

1) Training parallel corpus: EN→DE, EN→ES (WMT14) and EN→FR (WMT13)
2) Translation Test: WMT test dataset
3) Disambiguation Bias Challenge Sets: WSD Bias and Adversarial

Experiments (Disambiguation Bias Challenge Sets)

- WSD Bias

Contains most likely incorrectly translated words due to some co-occurrences

e.g. “a lot of money was spent to renovate the capital”

Likely translated word set: 资本 (sense of amount of money)

It quantifies the intrinsic bias (shortcut) the model learned during training.

- Adversarial

Original: “they met in the spring of 2020” -> sense of season

Corrupted: “they met in the hot spring of 2020” -> sense of water

Experiments

- Models

1) Translation model: 6 encoder layers + 6 encoder layers
2) (Pretrained) Sense embeddings: ARES [Scarlini et al., 2020b]
3) WSD model: EWISER

- Datasets

1) Training parallel corpus: EN→DE, EN→ES (WMT14) and EN→FR (WMT13)
2) Translation Test: WMT test dataset
3) Disambiguation Bias Challenge Sets: WSD Bias and Adversial
4) DIBIMT: correct or incorrect translation equivalents

DIBIMT

Figure 1: Example of an annotated dataset item. Target word is **shot**, in its meaning of a "small drink of liquor". We expect translations to contain, for example in Italian, *goccio* (lit. a drop), but not, for example in Spanish, *pistolero* (a person who shoots).

Experiments

- Models
- Datasets
- Comparison systems

(1) OPUS: the same architecture and parameter count but much more data

(2) Mbart-50: feature bigger underlying models

Results

	EN → DE		EN → FR	EN → ES
	WMT14	WMT19	WMT14	WMT13
OPUS†	27.58	39.39	39.93	35.00
MBart-50‡	25.60	35.80	36.12	29.50
Baseline	26.34	36.93	38.05	32.82
Baseline+SCR	**27.26**	**37.74**	**38.48**	**33.18**
Baseline+SCR$_{-KL}$	26.13	36.45	37.85	33.15
Baseline+SCR$_{-ARES}$	25.75	35.93	37.33	32.49
Baseline+SCR$_{-AR}$	26.11	36.74	37.38	32.93
Baseline+SCR$_{\text{RAND}}$	25.63	34.79	/	/

- Translation quality (BLEU)
- Baseline + SCR vs. Baseline/Mbart-50/OPUS
- Ablation: _KL (without SCR); _ARES (random sense embedding);

  _AR (annotation refinement); RAND (random sense)

Results

- Disambiguation ability
- Baseline+SCR vs. baseline+
- Baseline+SCR_AR ~ EWISER

MODEL	WSD Bias ↓	Adversarial ↓
Baseline	12.27	4.48
Baseline+SCR	**11.23**	**4.21**
Baseline+SCR$_{-KL}$	12.43	5.14
Baseline+SCR$_{-ARES}$	12.53	4.75
Baseline+SCR$_{-AR}$	13.07	4.93
Baseline+SCR$_{\text{RAND}}$	12.56	5.04
EWISER	13.70	/
Baseline$_{cf}$	11.86	/
Baseline+SCR$_{cf}$	9.91	/

Results

- Baseline vs. Baseline+SCR
- Baseline vs. (OPUS and Mbart-50)

Model	EN $\rightarrow$ DE	EN $\rightarrow$ ES
OPUS†	27.99	36.66
MBart-50‡	28.73	33.89
Baseline	24.00	26.44
Baseline+SCR	25.00	25.84

Table 4: Accuracy scores on DiBiMT. † and ‡ have the same meaning as in Table 1. Higher is better.

Results

Source sentence / Reference sentence / Baseline output / Enhanced output	Target sense	Wrong sense
S: [...] that both first words start with the same **letter**. **R**: [...] dass beide Begriffe mit demselben **Buchstaben** beginnen. **B**: [...] dass beide Wörter mit dem gleichen **Brief** beginnen. **E**: [...] dass beide Wörter mit dem gleichen **Buchstaben** beginnen.	*alphabet symbol*	*written message*
S: At least since the **fall** of 2008, leading economies' officials have agreed [...] **R**: Spätestens seit **Herbst** 2008 stimmen die Vertreter führender [...] **B**: Zumindest seit dem **Fall** 2008 haben sich die Beamten [...] **E**: Zumindest seit dem **Herbst** 2008 haben sich die Beamten [...]	*season*	*act of falling*
S: The construction of the Deurganck dock **lock** is [...] **R**: Der Bau der **Schleuse** am Deurganck-Dock ist [...] **B**: der Bau der Deurganck-**Hafensperre** ist [...] **E**: der Bau der Deurganck-**Hafenschleuse** ist [...]	*segment of a canal*	*blockade*

Table 3: Examples of sentences that were disambiguated correctly by our enhanced model but not by the baseline. Ambiguous word is in **blue**, wrong translation is in **red**, correct translation is in **green**.

- S: source sentence; R: reference sentence; B: baseline; E: enhanced (paper)

Conclusions

- NMT + WSD
- Annotated Corpus; Model (Input + loss function)
- Effectively reduce lexical ambiguation bias without losing quality

A Psycholinguistic Analysis of BERT's Representations of Compounds

Lars Buijtelaar
University of Amsterdam
lars.buijtelaar@student.uva.nl

Sandro Pezzelle
ILLC, University of Amsterdam
s.pezzelle@uva.nl

EACL 2023

EACL 2023

- 17th Conference of the European Chapter of the Association for Computational Linguistics
- 2023, 2-6 May, Dubrovnik
- Acceptance rate

(1) Main: 24.1% (281/1166)

(2) Findings: 17.2% (201/1166)

EACL – Lexical Semantics, Discourse and Anaphora

1. A Psycholinguistic Analysis of BERT's Representations of Compounds
2. A Systematic Search for Compound Semantics in Pretrained BERT Architectures
3. Bridging the Gap Between BableNet and HowNet: Unsupervised Sense Alignment and Sememe Prediction
4. What happens before and after: Multi-Event Commonsense in Event Coreference Resolution
5. A simple but effective model for attachment in discourse parsing with multi-task learning for relation labeling
6. Exploring Category Structure with Contextual Language Models and Lexical Semantic Networks

Introduction

- English Compounds, e.g., sunlight, bodyguard
- Lexicalization: as part of lexicon
- Compositionality: combine the meaning of its constituents
- What does the machine know if ”he” is capable of representing lexical (compound in our case) meaning? [Psycholinguistic]

1) which part does the compound rely more on?

E.g., *handgun* relies more on *gun* than *hand*

2) What is the degree of semantic transparency/compositionality?

E.g., *sunlight* is more semantically transparent than *muskrat* (麝鼠)

Introduction

- Transformers are shown to produce well-aligned word representation
- Lower layers
- But measured only in the word level, not in their parts

- Similar task: binary classification of literality of a compound
- But they need to train a binary classifier, not measure the embedding itself.

Contributions

- Measure (Lexeme Meaning Dominance) LMD and (Semantic Transparency) ST for BERT
- Calculate their correlation with human judgments
- Comprehensive experiments on model versions, contexts, pooling methods, layers

Related Work

- Compounds are one of the favorite subjects of psycholinguistic (and linguistics) research
- Compositionality, frequency, semantic transparency, or headedness (Gagné and Spalding, 2009; Marelli et al., 2009; Marelli and Luzzatti, 2012; Juhasz et al., 2015)
- Static embeddings and compositional models of distributional semantics
- Lack of contextualized representation by Transformer-based encoders
- BERTology: Lower layers: lexico-semantic; higher layers: contextualized

Data

- A psycholinguistic dataset with human judgements on compound LMD and ST
- LMD: lexeme meaning dominance
- ST: semantic transparency
- 629 lexicalized English compounds annotated by 189 participants

compound	LMD [0,10]	ST [1,7]
handgun	8.13 →	6.29 ↑
bodyguard	7.27 →	5.64 ↑
policeman	3.07 ←	6.13 ↑
wartime	3.47 ←	6.31 ↑
muskrat	7.53 →	2.80 ↓
primrose	7.93 →	2.00 ↓
milestone	3.36 ←	2.21 ↓
cheapskate	2.00 ←	2.00 ↓

Table 1: A few examples from the dataset with either high ↑ or low ↓ ST and either low ← or high → LMD. E.g., the meaning of *handgun* is deemed highly transparent and based more on the right than the left constituent.

Method

Models: BERT_base and BERT_large

Word-Level Representation

- No-Context (NC) -> without any context: <CLS>(snow, ##board)<SEP>

(1) nospec: emb(word)

(2) withcls: emb(word) + emb([CLS]))

(3) all: emb(word) + emb([CLS]) + emb([SEP])

- In-Context (C)

nospec setting and average the sampled occurrences in a corpora

Measures

- <compound/c, left, right>; L: cos(left, c); R: cos(right, c)
- LMD: $LMD(c) = 5(R - L) + 5$

  R=1 and L=0 -> LMD = 10; R=0 and L=1 -> LMD = 0
- ST:

$$ST(c) = \frac{6(L+R)}{2} + 1$$

L=R=1 -> ST=7; L=R=0 -> ST=1;

Different from LMD when R=1/0 and L=0/1

Evaluation

Difference from human judgements

- Mean absolute distance (MAE)
- Spearman correlation

Results

- **Baseline**: Glove
- **Trend**: increase with higher layer
- **Best**: Bert-large C (~0.6)
- C vs. NC
- Bert vs. Glove
- Lexico-semantics vs. contextualized

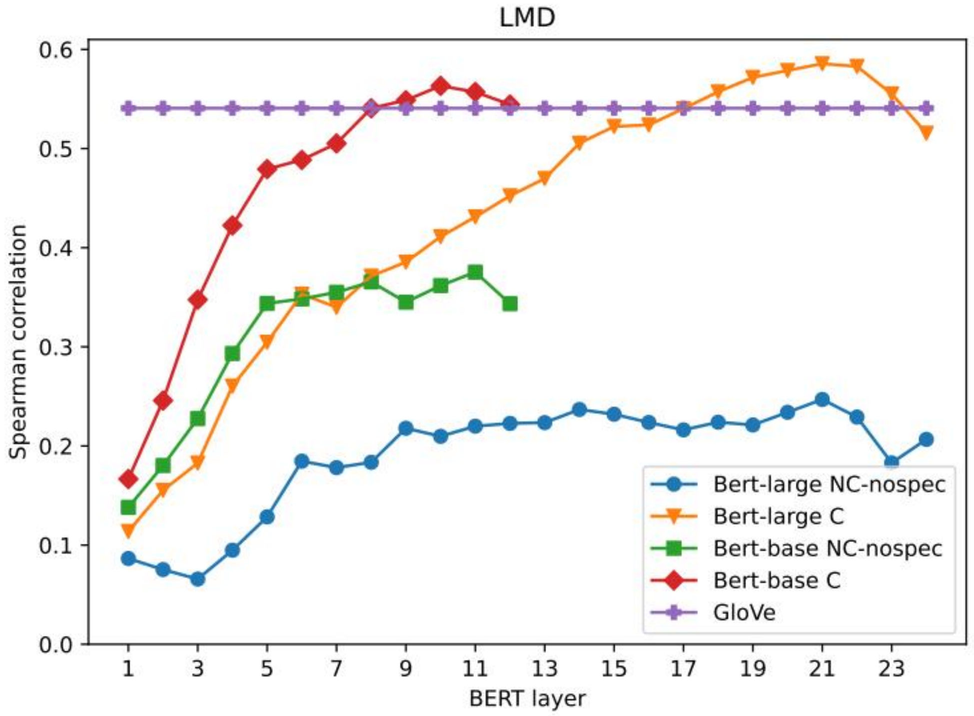


Figure 1: LMD. ρ against model layers. For out-of-context BERT models, we only report the best-performing nospec setting. Best viewed in color.

Results

model	setting	metric (best layer)	
		MAE ↓	Spearman ρ ↑
GloVe	–	2.657	0.304
$\text{BERT}_{\text{base}}$	`NC_nospec`	0.953 (6)	0.316 (5)
	`NC_all`	1.129 (10)	0.234 (1)
	`NC_withcls`	0.989 (1)	0.275 (3)
	`C`	*0.899 (9)*	*0.415 (9)*
$\text{BERT}_{\text{large}}$	`NC_nospec`	0.989 (9)	0.195 (6)
	`NC_all`	1.118 (24)	0.113 (1)
	`NC_withcls`	1.024 (6)	0.139 (1)
	`C`	**0.876 (19)**	**0.476 (20)**

Table 3: ST. Results in **bold** and *italic* are the best and second-best in the column, respectively. Results are from a model's best-performing layer (in parentheses).

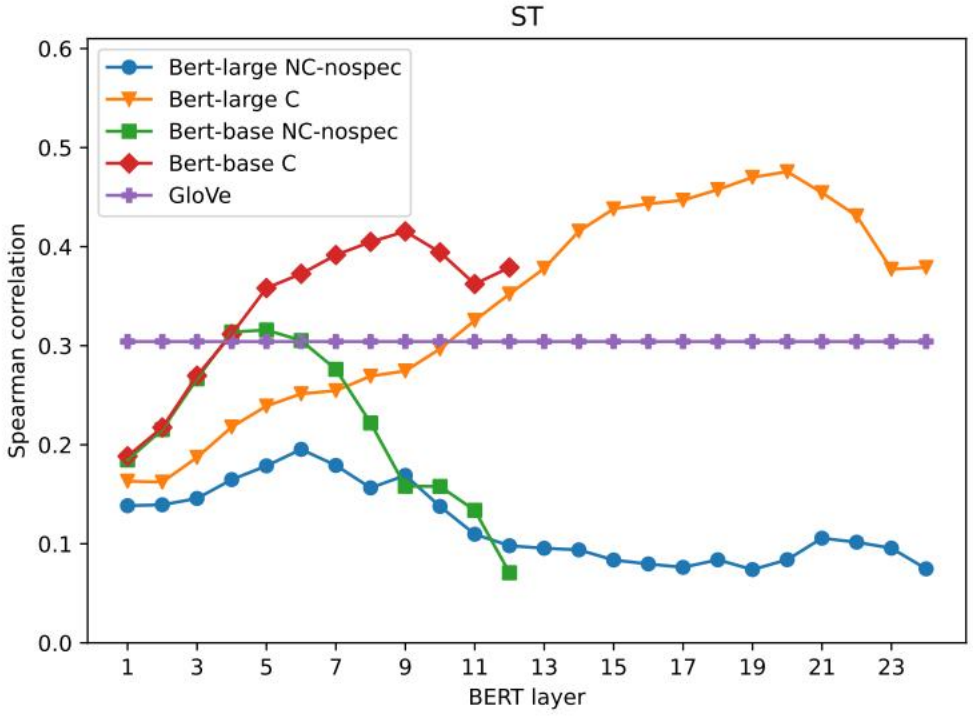


Figure 2: ST. ρ against model layers. For out-of-context BERT models, we only report the best-performing nospec setting. Best viewed in color.

- ST is more challenging than LMD (0.47
- BERTlarge vs. Glove (a larger gap)
- Different trends for NC setting in the right figure

Results

- LMD:

Muskrat (trend, higher layer)

- ST:

Milestone (different from LMD)

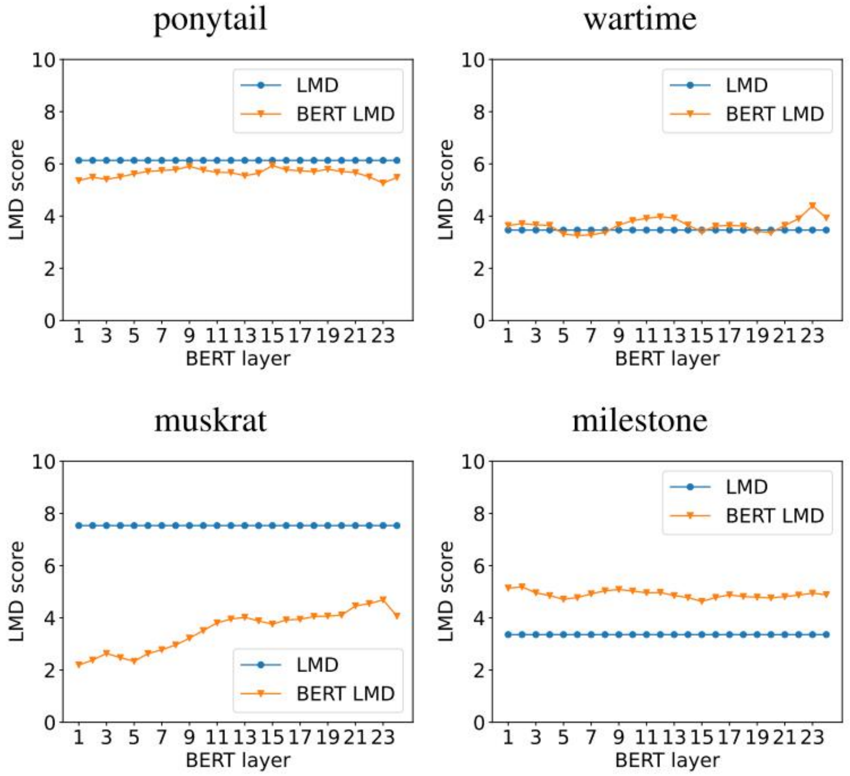


Figure 3: Examples where C BERT$_{large}$ is good (top) and bad (bottom) in approximating human LMD. From top left, clockwise: *ponytail*, *wartime*, *milestone*, *muskrat*.

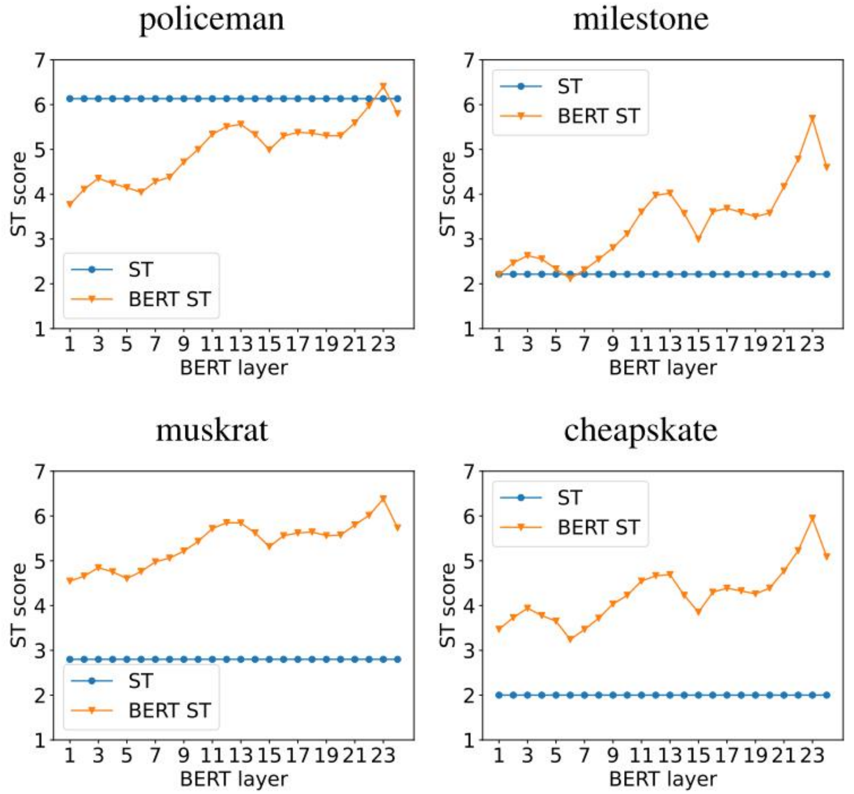


Figure 4: Examples where C BERT$_{large}$ is good (top) and bad (bottom) on ST. From top left, clockwise: *policeman*, *milestone*, *cheapskate*, *muskrat*.

Which factors drive the prediction?

- Linear regression model for LMD and ST
- Independent variables:

(1) the number of tokens

(2) the frequency of the compound in the dataset

(3) the compound/modifier/head concreteness

- (Results) LMD: the concreteness of the head and the modifier
- ST: number of tokens; compound and modifier concreteness

LMD: reversed compound

- wartime vs. timewar
- whether being no or little aware of the semantic and syntactic (modified and modifier) -> different LMD, thus different correlation
- Or just a correlation? -> similar LMD, thus similar correlation

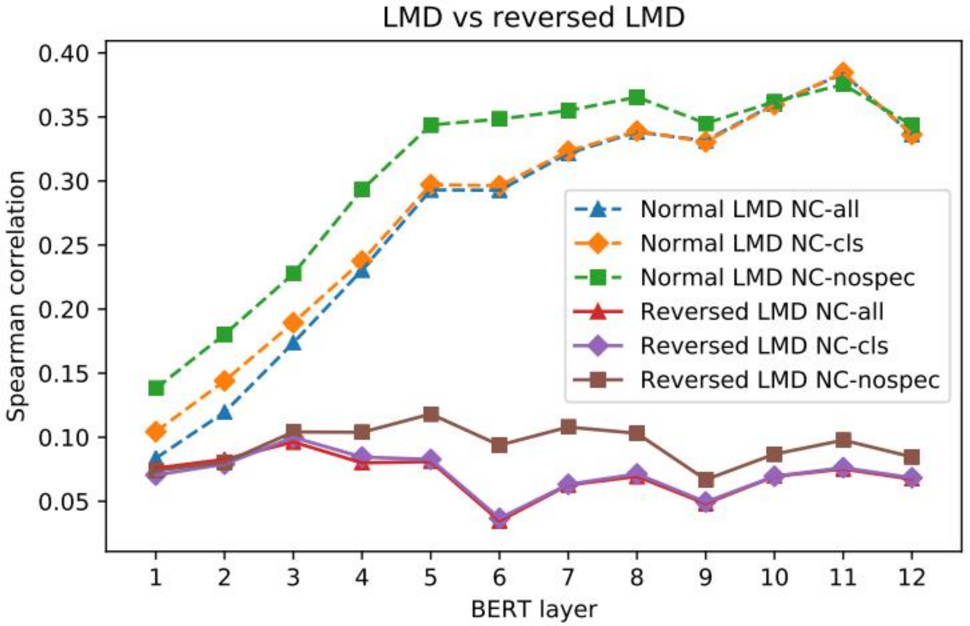


Figure 5: ρ for LMD vs LMD *reversed* values by out-of-context BERT$_{\text{base}}$ across layers. Best viewed in color.

ST: Weighted Constituents

$$ST(c) = \frac{6(L+R)}{2} + 1$$

- Weighted version for left and right
- Similar trend
- 0.5 performs well

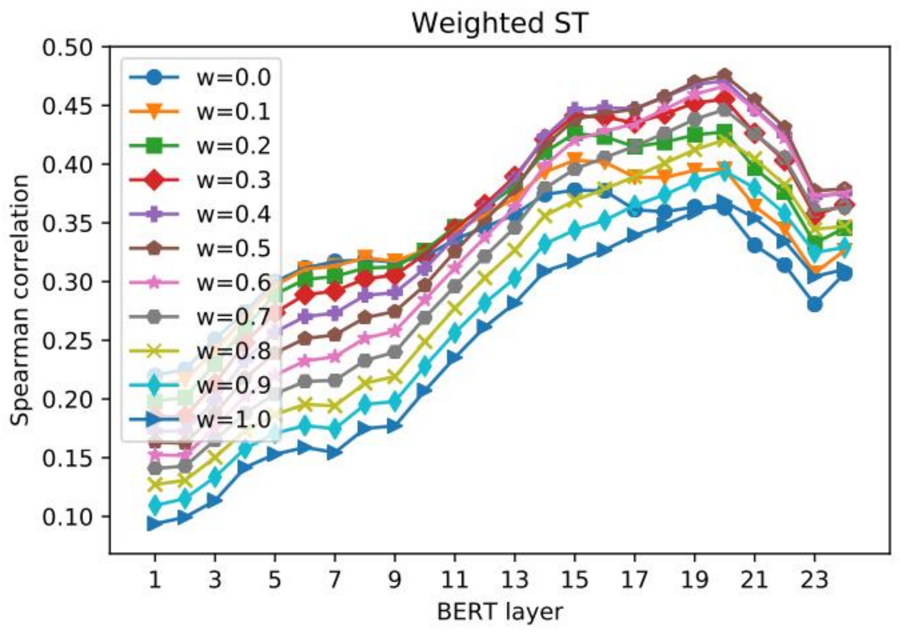


Figure 6: ρ for *weighted* ST by in-context BERT$_{\text{large}}$ across layers. Each weight stands for the weight assigned to the left constituent. Best viewed in color.

Conclusion

- How BERT represents the meaning of lexicalized compounds
- A psycholinguistic angle (LMD and ST)
- A specific, context-dependent representation

补充

- 复合构词是汉语中常见的合成方式（重叠和附加为次要）
- 词库词vs.词法词（葡萄<文化<蝴蝶<学校<鸡肉）
- 前者语义更加特异、规则更加晦涩、构造理据更加不明；后者相反

Q & A

THANK YOU

EWISER

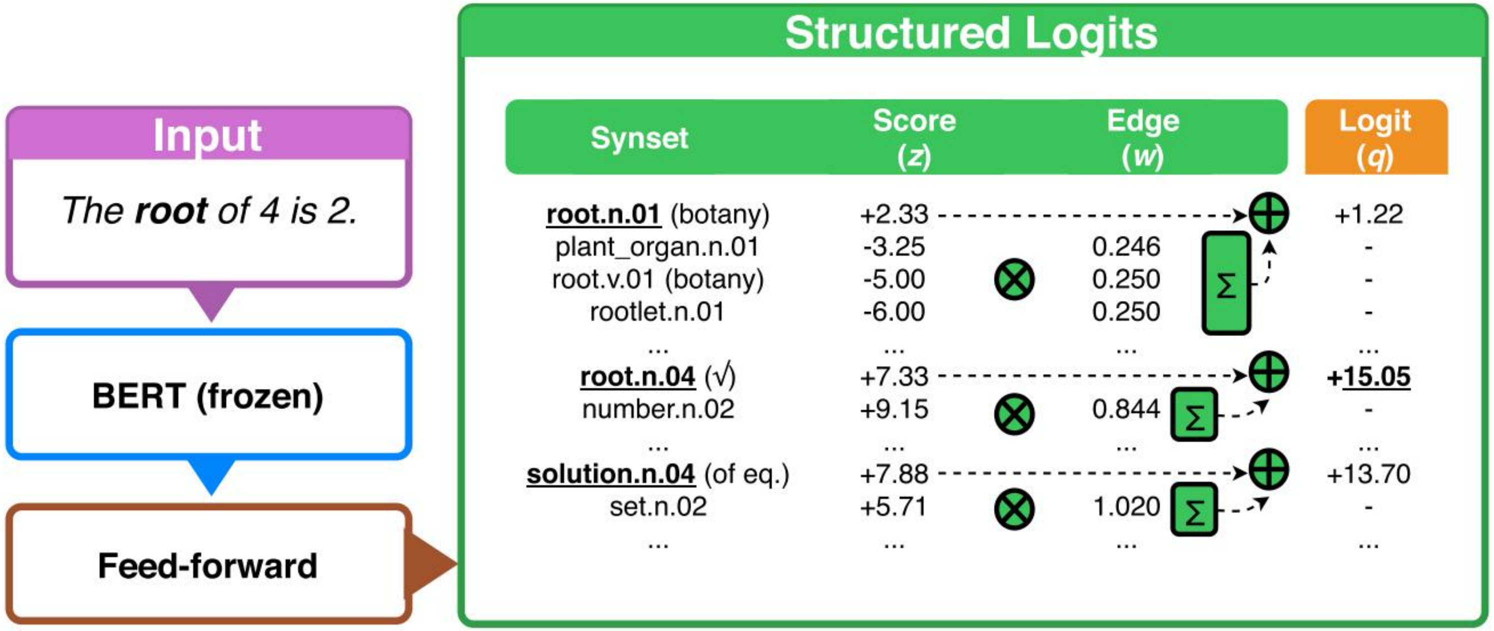


Figure 1: The structured logits mechanism in EWISER. The example input is the sentence "The *root* of 4 is 2." Scores for a selection of synsets representing possible senses of *root* are shown. Going from left to right, the "hidden" logits (**z**) of related synsets are multiplied by the edge weights, summed together, and then added to the "hidden" logits of the related synsets, resulting in the "final" logits (**q**).

BabelNet

- (multilingual) Synset-based
- Even Mutimodal!
- Coordinate with Wordnet
- Not every concept maps to every language.