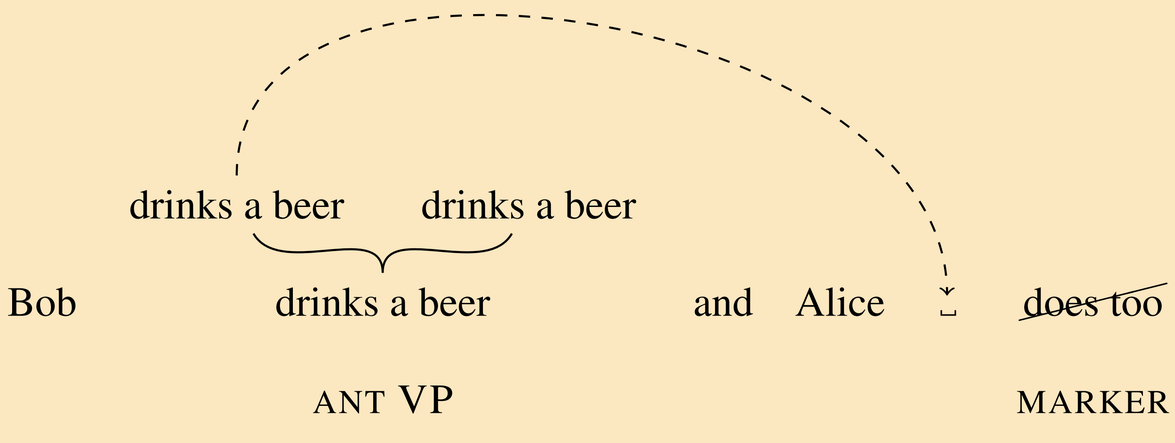
Evaluating Composition Models for VP-Elliptical Sentence Embeddings

Gijs Wijnholds and Mehrnoosh Sadrzadeh
g.j.wijnholds@qmul.ac.uk mehrnoosh.sadrzadeh@qmul.ac.uk

VERB PHRASE ELLIPSIS

part of a sentence is missing but can be recovered from surrounding context

Problem: textual similarity datasets do not cover ellipsis.

Hypothesis: resolving ellipsis benefits sentence embedding quality.

Method: we built two datasets specifically for ellipsis, a disambiguation and a similarity task. We provide comprehensive evaluation, comparing composition models on multiple vector spaces with state of the art sentence encoders.

VP - ELLIPTICAL SENTENCE EMBEDDINGS

Higher-Order Unification (Dalrymple et al. 1991)

(b_1) $chase(cats, dogs) \wedge P(children)$

(b_2) $P = \lambda x.chase(x, dogs)$

(b_3) $(b_1) \rightsquigarrow_\beta chase(cats, dogs) \wedge chase(children, dogs)$

Lambdas and Vectors (Muskens/Sadrzadeh 2016)

$\cdot$	:=	$\lambda rvi.r \cdot v_i$	$: RVV$
$+$	:=	$\lambda vwi.v_i + w_i$	$: VVV$
$\odot$	:=	$\lambda vwi.v_i \cdot w_i$	$: VVV$
$\times_1$	:=	$\lambda mvij. \sum_j m_{ij} \cdot v_j$	$: MVV$
$\times_2$	:=	$\lambda cvijk. \sum_k c_{ijk} \cdot v_k$	$: T^3VM$

c	$\mathcal{H}(c)$	$\mathcal{T}(c)$
cn	**cn**	V
tv	$\lambda uv.(\mathbf{tv} \times_2 v) \times_1 u$	VVV
coord	$\lambda P.\lambda Q.P \nabla Q$	VVV

Tensor-Based Modelling of VP-Ellipsis:

$$((\mathbf{chase} \times_2 \mathbf{dogs}) \times_1 \mathbf{cats}) \nabla ((\mathbf{chase} \times_2 \mathbf{dogs}) \times_1 \mathbf{children})$$

EVALUATION SETUP

New Verb Disambiguation Dataset

$\langle$ *man draw sword, man depict sword* $\rangle$
$\langle$ *man draw sword, man attract sword* $\rangle$

↓

man depict sword and artist does too
man attract sword and artist does too

New Sentence Similarity Dataset

$\langle$ *school encourage child, employee leave company* $\rangle$

↓

$\langle$ *school encourage child and parent does too, employee leave company and student does too* $\rangle$

Vector Spaces

Custom trained on UKWaCkypedia (ca. 123M sentences)

CB: count based, PPMI, D=2000
W2V: SGNS space, D=300
GloVe: D=300
FT: FastText space, D=300

Verb Tensors

Relational

$$\overline{verb} = \sum_i subj_i \otimes obj_i$$

Kronecker

$$\widetilde{verb} = \overrightarrow{verb} \otimes \overrightarrow{verb}$$

RESULTS

Model Type	Embedding
Linear Vector	$\overrightarrow{subj} \star \overrightarrow{verb} \star \overrightarrow{obj} \star \overrightarrow{and} \star \overrightarrow{subj^*} \star \overrightarrow{does} \star \overrightarrow{too}$
Non-Linear Vector	$\overrightarrow{subj} \star \overrightarrow{verb} \star \overrightarrow{obj} \star \overrightarrow{subj^*} \star \overrightarrow{verb} \star \overrightarrow{obj}$
Tensor-Based	$T(\overrightarrow{subj}, \overrightarrow{verb}, \overrightarrow{obj}) \star T(\overrightarrow{subj^*}, \overrightarrow{verb}, \overrightarrow{obj})$

Composition

$\star$ = addition or multiplication

Base: "*subj verb obj and subj* does too*"
Res: "*subj verb obj and subj* verb obj*"
Abl: "*subj verb obj subj**"

Encoders

Pretrained Encoders
D2V: Doc2Vec, D=300
ST: Skip-Thoughts, D=4800
IS: InferSent, D=4096
USE: Universal Sentence Encoder, D=512

	CB	W2V	GloVe	FT
Verb Only Vector	.4363	.2406	.4451	.2290
Verb Only Tensor	.3295	.4376	.3942	.3876
Add. Linear	.4416	.2728	.3046	.1409
Mult. Linear	3250	-.0123	.1821	.2928
Add. Non-Linear	.4448	.3275	.3262	.1399
Mult. Non-Linear	.5029	.2087	.2446	.0440
Best Tensor	**.5385**	.4621	.3688	.4937
2nd Best Tensor	.5263	.4544	.3581	.4652

Verb Disambiguation

	D2V1	D2V2	ST	IS1	IS2	USE
Base	.1448	.2432	-.1932	.3471	**.3841**	.2693
Res	.2340	.2980	-.1720	.3436	.3373	.2770
Abl	.1899	.2423	-.1297	.3525	.3571	.2402

	CB	W2V	GloVe	FT
Verb Only Vector	.4562	.5833	.4348	.6513
Verb Only Tensor	.3946	.5664	.4426	.5337
Add. Linear	.7000	.7258	.6964	.7408
Mult. Linear	.6330	.1302	.3666	.1995
Add. Non-Linear	.6808	**.7617**	.7103	.7387
Mult. Non-Linear	.7237	.3550	.2439	.4500
Best Tensor	.7410	.7061	.4907	.6989
2nd Best Tensor	.7370	.6713	.4819	.6871

Sentence Similarity

	D2V1	D2V2	ST	IS1	IS2	USE
Base	.5901	.6188	.5851	.7785	.7009	.6463
Res	.6878	.6875	.6039	**.8022**	.7486	.6791
Abl	.1840	.6599	.4715	.7815	.7301	.6397

Datasets, code, and custom vector/tensor spaces all available online:
github.com/gijswijnholds/compdisteval-ellipsis

gijswijnholds.github.io

Queen Mary University of London