香港中文大學
The Chinese University of Hong Kong

Functional Distributional Semantics at Scale

Chun Hei Lo[1] Hong Cheng[1] Wai Lam[1] Guy Emerson[2]

[1]The Chinese University of Hong Kong [2]University of Cambridge

UNIVERSITY OF CAMBRIDGE

Summary

Functional Distributional Semantics (FDS) models lexical and sentence-level semantics with functions using distributional information. Previous implementations of FDS focus on subject–verb–object (SVO) triples only. We devise computationally efficient and linguistically motivated methods for applying FDS to arbitrary sentences.

Functional Distributional Semantics

Core idea. A sentence refers to a set of *entities*, and a word is a *predicate* that is true or false of entities. To generalize, an entity is represented as a pixie, and a predicate is a *semantic function* that maps pixies to probability of truth. The generative model of FDS is illustrated in Fig. 1.

Model Learning. Given the observed predicates R and argument structure A of a *Dependency Minimal Recursion Semantics (DMRS)* graph (see Fig. 1 for an example graph), $\max P(R \mid A)$.

Linguistic Challenges. Moving away from SVO triples means the semantics of adverbs, adjectives, adpositions, conjunctions, and quantifiers need to be addressed. Moreover, the undirected graphical model is thus more unsuitable for predicate-specific interpretations of argument roles (see Table 1).

Computational Challenges. Computing the prior of pixies is intractable in CaRBM (see Table 1). An alternative proposal of adopting a Gaussian Markov Random Field scales to $\mathcal{O}(d^3n^3)$ time, which is prohibitive for larger graphs.

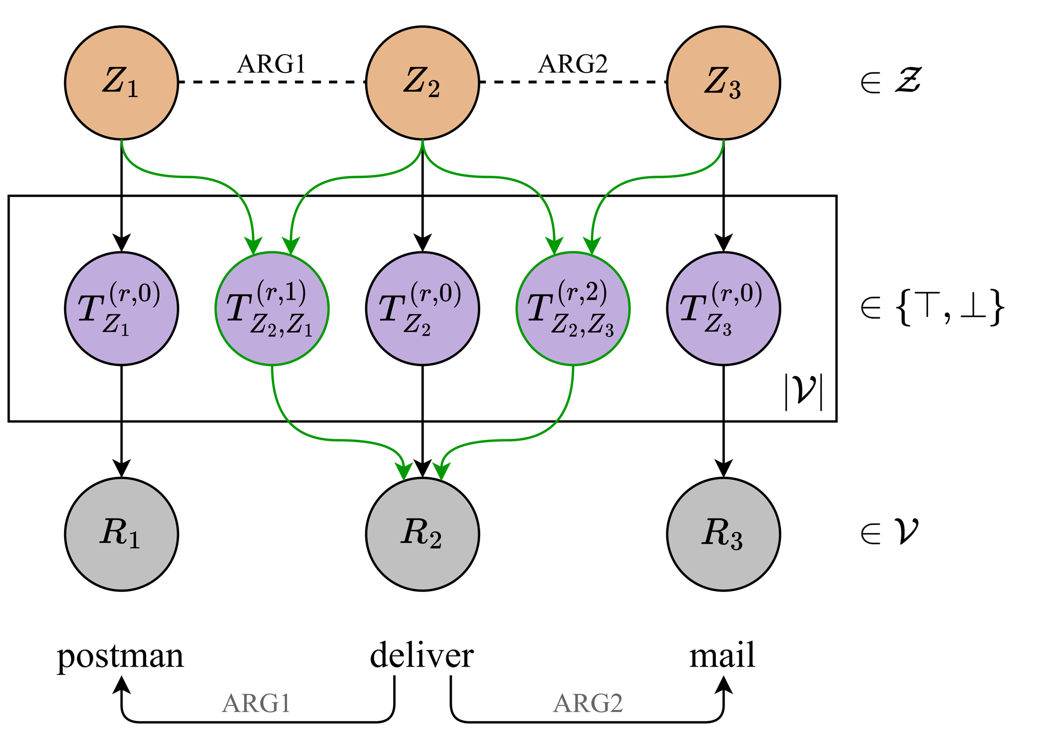

Figure 1. Top: probabilistic graphical model that generates the sentence 'postman deliver mail'; bottom: the simplified DMRS graph of the sentence, where R_1=*postman*, R_2=*deliver*, R_3=*mail* and $A = \{(2, 1, \text{ARG1}), (2, 3, \text{ARG2})\}$. Argument information only contributes to the world model in previous implementations (in dashed lines); we propose that it is used only in the lexical model (in green lines) (See Table 1).

Enriching the Lexical Model

Neo-Davidsonian Event Semantics. Different types of modifications, e.g., adverbial modification, can be handled by introducing event arguments:

$$\text{deliver}(e_1) \wedge \text{ARG1}(e_1, x) \wedge \text{ARG2}(e_1, y) \wedge \underline{\text{quick}(e_2, e_1)}$$

Semantic Functions. On top of the unary functions in (1), we add binary ones in (2):

$$P\left(T^{(r,0)}_{Z_e}{=}\top \;\middle|\; z_e\right) = t^{(r,0)}(z_e) \tag{1}$$

$$P\left(T^{(r,a)}_{Z_e,Z_x}{=}\top \;\middle|\; z_e, z_x\right) = t^{(r,a)}(z_e, z_x) \tag{2}$$

This way, dropped arguments (see (3)) as well as adverbs (and adjectives) and conjunctions (see Fig. 2) are handled naturally. Table 1 shows a summary of the changes.

$$P\left(T^{(r,0)}_{Z_e} \wedge T^{(r,2)}_{Z_e,Z_y}{=}\top \;\middle|\; z_e, z_y\right) = t^{(r,0)}(z_e)t^{(r,2)}(z_e, z_y) \tag{3}$$

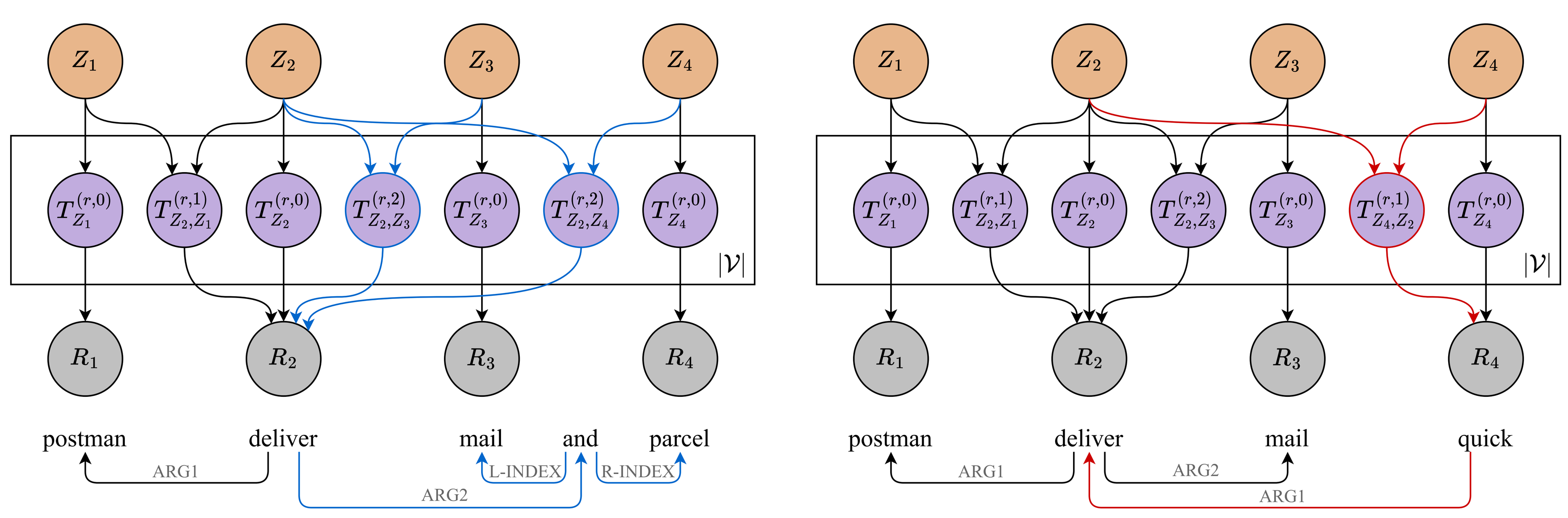

Figure 2. Extending the DMRS in Fig. 1 with adverbs and conjunctions.

Previous FDS	Our Proposal
$\mathcal{Z} = \{0,1\}^d$	$\mathcal{Z} = \mathbb{R}^d$
$P(z \mid A) \propto \exp\left(\sum_{(i,j,a)\in A} {z_i}^\top W^{(a)} z_j\right)$ (CaRBM) $P(r_i \mid z) \propto t^{(r_i,0)}(z_i) \quad \forall i$	$P(r_i \mid z) \propto t^{(r_i,0)}(z_i) \quad i \in \{1,3\}$ $P(r_2 \mid z) \propto t^{(r_2,0)}(z_2)t^{(r_2,1)}(z_2, z_1)t^{(r_2,2)}(z_2, z_3)$
$\mathcal{V} \subseteq$ nouns and verbs	$\mathcal{V} \subseteq$ nouns, verbs, adjectives, and adverbs

Table 1. Comparison between previous and our proposed formulation for the DMRS in Fig. 1.

Variational Autoencoder

Probabilistic Encoder. Given an observed DMRS graph with n pixies, the approximate posterior is given by:

$$q_\phi(z \mid R, A) = \prod_{i=1}^{n} \mathcal{N}(z_i; \mu_{Z_i}, \sigma^2_{Z_i} I) \tag{4}$$

For each pixie Z_i, the mean μ_{Z_i} and log variance $\ln \sigma^2_{Z_i}$ are inferred (f can be the identity function or $\tanh$):

$$h^{(Z_i)} = f\left(\frac{1}{n}\sum_{j=1}^{n} e^{(r_j, a_{j,i})}\right) \tag{5}$$

$$\mu_{Z_i} = W^\top h^{(Z_i)} + c_1 \tag{6}$$

$$\ln \sigma^2_{Z_i} = w^\top h^{(Z_i)} + c_2 \tag{7}$$

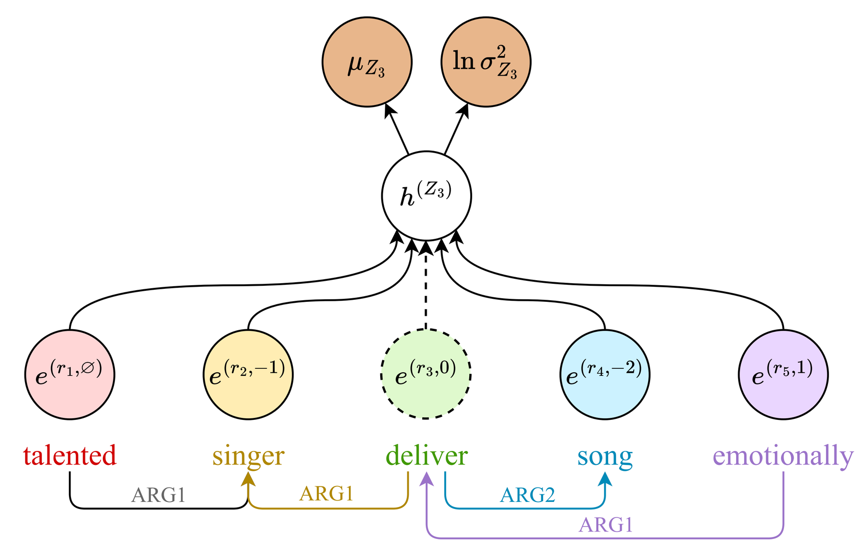

Figure 3. An encoder for inferring the posterior distribution of the pixie of *deliver* in '*talented singer deliver song emotionally*'. Dropout is applied to prevent learning shortcuts (in dashed lines).

Probabilistic Decoder. Given the inferred posterior $q_\phi(z \mid R, A)$, we compute the probabilities of truth of predicates over the pixie distribution. Linear classifiers in (8) and (9) allow probit approximation in (10) for the expectation of $t^{(r,0)}(z_i)$, W.L.O.G. for $t^{(r,a)}(z_i, z_j)$ (S is the sigmoid function and $z_{i,j}$ is the concatenation of z_i and z_j).

$$t^{(r,0)}(z_i) = S\left({v^{(r,0)}}^\top z_i + b^{(r,0)}\right) \tag{8}$$

$$t^{(r,a)}(z_i, z_j) = S\left({v^{(r,a)}}^\top z_{i,j} + b^{(r,a)}\right) \tag{9}$$

$$\mathbb{E}_{q_\phi}\left[t^{(r,0)}(z_i)\right] \approx S\left(\frac{{v^{(r,0)}}^\top \mu_{Z_i} + b^{(r,0)}}{(1 + \frac{\pi}{8}\sigma^2_{Z_i})^{\frac{1}{2}}}\right) \tag{10}$$

Final Objective. For each observed predicate r_i, we sample K negative predicates $N(i)$, assuming them to be false of the inferred pixies. Reformulated the β-VAE with variance regularization, we maximize (11).

$$\tilde{\mathcal{L}}_{\phi,\theta}(R \mid A) = \sum_{i=1}^{n} \mathcal{C}_i + \sum_{(i,j,a)\in A} \mathcal{C}_{i,j,a} - \beta\frac{d}{2}\sum_{i=1}^{n}\left(\sigma^2_{Z_i} - \ln \sigma^2_{Z_i}\right) \tag{11}$$

$$\text{where} \quad \mathcal{C}_i = \ln \mathbb{E}_{q_\phi}\left[t^{(r_i,0)}(z_i)\right] + \sum_{r' \in N(i)} \ln \mathbb{E}_{q_\phi}\left[1 - t^{(r',0)}(z_i)\right],$$

$$\mathcal{C}_{i,j,a} = \ln \mathbb{E}_{q_\phi}\left[t^{(r_i,a)}(z_i, z_j)\right] + \sum_{r' \in N(i)} \ln \mathbb{E}_{q_\phi}\left[1 - t^{(r',a)}(z_i, z_j)\right]$$

Experiments

Models Training

Data Set. *Wikiwoods*: 36m sentence–DMRS pairs (254m tokens) after preprocessing.

Tuning. Each of our models is tuned on the development set of RELPRON (described below), and have their outputs averaged over three random seeds.

Evaluation on Semantic Composition

Data Set. *RELPRON*: Retrieve the corresponding properties for each term.

Term (noun)	Property (subj./obj. relative clause)
watch	device that astronomer use
telescope	device that keep time
observatory	device that observatory have
license	building that astronomer own
...	organization that army install
	...

Table 2. Example instances in dev. set of RELPRON. Underlined is a confounding pair with lexical overlap.

Model	MAP
Pixie Autoencoder (FDS)	0.19
Ensemble of PixieAE & vector add. (FDS)	0.49
$\text{BERT}_{\text{BASE}}$ (tuned; with full stop)	0.67
$\text{BERT}_{\text{BASE}}$ (tuned; without full stop)	0.20
$\text{FDSAS}_{\text{tanh}}$	0.48
FDSAS_{id}	0.58

Table 3. Results on test set of RELPRON.

Evaluation on Verb Disambiguation

Data Sets. *GS2011* and *GS2013*: For each SVO–landmark pair, rate the semantic similarity of the verb in the SVO and the landmark verb; *GS2012*: With adjectives.

Adj-Subject-Verb-Adj-Object	Landmark (verb)	Similarity Annotations (1–7)
social service meet educational need	visit	1, 2, ...
social service meet educational need	satisfy	7, 6, ...
young boy meet little girl	visit	3, 2, ...
small child write single word	spell	6, 7, ...
local people write open letter	spell	2, 3, ...
...	...	...

Table 4. Example instances of GS2011 (GS2012, with the grey adjectives).

Model	ρ
Joint Learning of Phrase Embeddings (Ensemble)	0.52
Pixie Autoencoder (FDS)	0.41
$\text{BERT}_{\text{BASE}}$ (Baseline)	0.39
$\text{FDSAS}_{\text{tanh}}$	0.44
FDSAS_{id}	0.44
Inter-annotator agreement	0.58

Table 5. Results on GS2011.

Model	ρ
Kronecker Model	0.26
Role-Filler Averaging Model with Residual Learning	0.37
$\text{BERT}_{\text{BASE}}$ (Baseline)	0.43
$\text{FDSAS}_{\text{tanh}}$	0.44
FDSAS_{id}	0.46
Inter-annotator agreement	0.59

Table 6. Results on GS2013.

Model	ρ
Dependency-based Compositional Semantics	0.33
Practical Lexical Function Model	0.36
$\text{BERT}_{\text{BASE}}$ (Baseline)	0.40
$\text{FDSAS}_{\text{tanh}}$	0.44
FDSAS_{id}	0.45
Inter-annotator agreement	0.46

Table 7. Results on GS2012.

chlo@se.cuhk.edu.hk