A Sense-Topic Model for Word Sense Induction with Unsupervised Data Enrichment

Jing Wang[1], Mohit Bansal[2], Kevin Gimpel[2], Brian Ziebart[1], Clement Yu[1]

[1]University of Illinois at Chicago

[2]Toyota Technological Institute at Chicago

Word Sense Induction

"…***images*** of him with his tailored arms across the shoulders of other leaders"

"…names such as Cottonelle trade on the ***image*** of cotton"

"…an impossibly heroic ***image***"

"It's an ***image*** that's as likely to make you feel icky"

"I consulted my own ***image*** expert"

"In the ***image*** of correlation matrix"

"…one being created in the ***image*** of God"

What are the meanings of each ***image*** ?

Task: automatically discover the possible senses of an ambiguous word in a given corpus.

- unsupervised
- senses represented by token clusters
- multiple senses per instance

Topic vs Sense

They are related, but distinct linguistic phenomena:

- topics are expressed by *all* word tokens in the document
- senses relate to a single ambiguous word and are expressed through the *local* context of that word
- Knowing the topic helps narrow down the set of plausible senses:

  "His reaction to the experiment was cold."

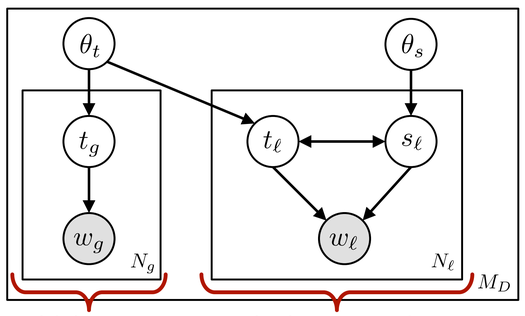
  (cold temperature, cold sensation, common cold, negative emotional reaction)

  If the topic is about the effect of low temperatures on physical health?

  (cold temperature, cold sensation, common cold, ~~negative emotional reaction~~)
- Knowing the sense can also help determine possible topics

  e.g. sense: cold ischemia ⟶ most probable topic: organ transplantation

Sense-Topic Model



Generating global context words:

$$\Pr(w_g|d,\theta_t,\psi_t)=\sum_{j=1}^{T}P_{\psi_{t_j}}(w_g|t_g=j)P_{\theta_t}(t_g=j|d)$$

Generating local context words (in window: 10 words before and 10 words after):

$$\Pr(w_\ell|d)=\sum_{j=1}^{T}\sum_{k=1}^{S}\Pr(w_\ell|t_\ell=j,s_\ell=k)\Pr(t_\ell=j,s_\ell=k|d)$$

Take inspiration from dependency networks (Heckerman et al., 2001):

$$\Pr(t_\ell=j,s_\ell=k|d)=\frac{1}{Z_d}\Pr(s_\ell=k|d,t_\ell=j)\Pr(t_\ell=j|d,s_\ell=k)$$

By introducing deficient modeling (Brown et al., 1993):

$$\Pr(s_\ell=k|d,t_\ell=j)=\frac{P_{\theta_s}(s_\ell=k|d)P_{\theta_{s|t_j}}(s_\ell=k|t_\ell=j)P_{\theta_{st}}(t_\ell=j,s_\ell=k)}{Z_{d,t_j}}$$

$$\Pr(t_\ell=j|d,s_\ell=k)=\frac{P_{\theta_t}(t_\ell=j|d)P_{\theta_{t|s_k}}(t_\ell=j|s_\ell=k)}{Z_{d,s_k}}$$

$$\Pr(w_\ell|t_\ell=j,s_\ell=k)=\frac{P_{\psi_{t_j}}(w_\ell|t_\ell=j)P_{\psi_{s_k}}(w_\ell|s_\ell=k)}{Z_{t_j,s_k}}$$

Inference

We use collapsed Gibbs sampling for posterior inference. The conditional posterior over topics for a global context word token can be computed by:

$$\Pr(t_g^{(i)}=j|d,t^{-i},\boldsymbol{s},\cdot)\propto\underbrace{\frac{C_{dj}^{DT}+\alpha}{\sum_{k=1}^{T}C_{dk}^{DT}+T\alpha}}_{\Pr(t=j|d,t^{-i},\boldsymbol{s},\cdot)}\underbrace{\frac{C_{ij}^{WT}+\alpha}{\sum_{k'=1}^{W_t}C_{k'j}^{WT}+W_t\alpha}}_{\Pr(w_g^{(i)}|t=j,t^{-i},\boldsymbol{s},\cdot)}$$

Unsupervised Data Enrichment

- **Adding context**

  i. Add the actual context from the corpus from which it was extracted

  ii. Find a semantically-similar(by using word embeddings (Mikolov et al., 2013)) sentence from an external corpus and append it to the instance as additional context

  Examples:

  ① Nigeria then sent troops to challenge the coup, evidently to restore the president and repair Nigeria's corrupt ***image*** abroad. *(image%1:07:01::/4)*

  ② When asked about the Bible's literal account of creation, as opposed to the attractive concept of divine creation, every major Republican presidential candidate —even Bauer—has squirmed, ducked, and tried to steer the discussion back to "faith," "morals," and the general idea that humans "were created in the ***image*** of God." *(image%1:06:00::/2 imag\%1:09:02::/4)*

  ③ I have recently deliberately begun to use variations of "kick ass" and "bites X in the ass" because they are colorful, evocative phrases; because, thanks to South Park, ass references are newly familiar and hilarious and because they don't evoke particularly vivid mental ***image*** of asses any longer. *(image%1:09:00::/4)*

  ④ Also, playing video games that require rapid mental rotation of visual ***image*** enhances the spatial test scores of boys and girls alike. *(image\%1:06:00::/4)*

  ⑤ Practicing and solidifying modes of representation, Piaget emphasized, make it possible for the child to free thought from the here and now; create larger images of reality that take into account past, present, and future; and transform those ***image*** mentally in the service of logical thinking. *(image\%1:09:00::/4)*

- **Adding instances** that are semantically similar to the instances in our original dataset. (Increase the number of instances in the dataset)

- **Weighting by word similarity**:

  Replicate each local context word according to its exponentiated cosine similarity to the target word.

Sense	Top-5 terms per sense
Sense-Topic Model	
1	include, depict, party, paint, visual
2	zero, manage, company, culture, figure
3	create, clinton, people, american, popular
+weight by similarity	
1	depict, create, culture, mental, include
2	picture, visual, pictorial, matrix, movie
3	public, means, view, american, story

Experiments

SemEval-2013 Dataset: includes 50 target ambiguous words, 4,644 instances.

	Model	Data Enrichment	Fuzzy B-cubed %	Fuzzy NMI %	AVG
1	1 sense	-	62.3	0	-
2	All distinct	-	0	7.09	-
3	unimelb	add 50k instances	48.3	6.0	17.02
4	AI-KU	add 20k instances	39.0	6.5	15.92
5	LDA(local)	none	47.1	5.93	16.71
6	LDA(full)	none	47.3	5.79	16.55
7	LDA(full)	add actual context	43.5	6.41	16.7
8	word embedding product	none	33.3	7.24	15.53
THIS PAPER					
9	Sense-Topic Model	none	53.5	6.96	19.3
10		add ukWac context	54.5	**9.74**	23.04
11		add actual context	**59.1**	9.39	**23.56**
12		add instances	58.9	6.01	18.81
13		weight by sim.	55.4	7.14	19.89

Bidirectionality Analysis

to measure the impact of the bidirectional dependency between the topic and the sense variables:

Model	B-cubed(%)	NMI(%)	AVG
Drop s⟶t	52.1	6.84	18.88
Drop t⟶s	51.1	6.78	18.61
Full	53.5	6.96	19.30

References

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. 2001. Dependency networks for inference, collaborative filtering, and data visualization. J. Mach. Learn. Res., 1:49-75.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Lingustics, 19(2): 263-311.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In Proc. of ICLR.