Introducing two Vietnamese Datasets for Evaluating Semantic Models of (Dis-)Similarity and Relatedness

Kim Anh Nguyen, Sabine Schulte im Walde, Ngoc Thang Vu
Institute for Natural Language Processing - University of Stuttgart, Germany
{nguyenkh,schulte,thangvu}@ims.uni-stuttgart.de

Motivation

- Verification of semantic models requires adequate gold-standard resources.
- There is still a lack of gold resources for low-resource languages.
- How do semantic models for English transfer to Vietnamese?

Annotation Procedure

- 200 annotators rate 400 word pairs.
- Each annotator rates 30 word pairs.
- Each word pair is rated by 15 annotators.

Annotation Task

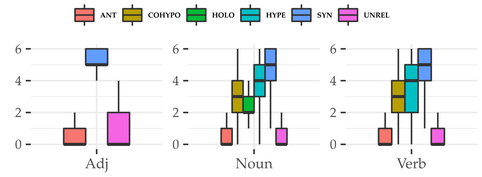
- Complete checkpoint question.
- Label kind of relation.
- Score degree of similarity on a 0-6 scale.

Contributions

1. Introduce two novel datasets for Vietnamese:
   - **ViCon**, a dataset of lexical contrast pairs.
   - **ViSim-400**, a rated dataset of semantic relation pairs.
2. Verify two datasets through standard and neural models.

Agreement in ViSim-400

Measures	*All*	Noun	Verb	Adjective
IAA-Mean ρ	0.86	0.86	0.86	0.78
IAA-Pairwise ρ	0.79	0.76	0.78	0.75
Krippendorff's α	0.78	0.76	0.78	0.86
STD	0.87	0.87	0.90	0.82



Criteria

- Similarity: synonymy (đội-*team* / nhóm-*group*), co-hyponymy (ô_tô-*car* / xe_đạp-*bike*)
- Relatedness: hypernymy, meronymy, functional association, antonymy (nóng-*hot* / lạnh-*cold*)
- Part-of-Speech: noun, verb, adjective
- Similarity (synonymy) vs. Dissimilarity (antonymy)

Concepts in ViCon

- Word pairs were drawn from Vietnamese Computational Lexicon (VCL).
- Extract all antonym and synonym pairs.
- Select randomly word pairs:
  - 600 adjective pairs
  - 400 noun pairs
  - 400 verb pairs
- In each word class, balance for:
  - number of antonymous and synonymous pairs
  - size of morphological classes

Concepts in ViSim-400

- Word pairs were drawn from VCL and Vietnamese WordNet (VWN).
- Extract all word pairs according to 5 semantic relations: synonymy, antonymy, hypernymy, co-hyponymy, holonymy
- Sample word pairs:
  - 200 noun pairs (6 relations)
  - 150 verb pairs (5 relations)
  - 50 adjective pairs (3 relations)
- In each word class, balance for:
  - number of word pairs w.r.t relations
  - size of morphological classes
  - number of lexical categories

Verification of ViSim-400

- Adopt a comparison of neural models on SimLex-999.
- Verify with three models:
  1. Skip-gram with Negative Sampling (SGNS)
  2. distributional Lexical Contrast Embeddings (dLCE)
  3. multitask Lexical Contrast Model (mLCM)
- Compute cosine similarity of all pairs.
- Use area under curve to distinguish between antonyms and synonyms.

Dataset	SGNS	mLCM	dLCE
ViSim-400	0.37	0.60	**0.62**
SimLex-999	0.38	0.51	**0.59**

Dataset	Model	Noun	Verb	Adj
ViSim400	SGNS	0.66	0.63	0.70
	mLCM	0.81	0.92	0.96
	dLCE	**0.92**	**0.95**	**0.98**
SimLex-999	SGNS	0.66	0.65	0.64
	mLCM	0.69	0.71	0.85
	dLCE	**0.72**	**0.81**	**0.90**

Verification of ViCon

- Verify with three co-occurrence models:
  1. positive pointwise mutual information (PPMI)
  2. positive local mutual information (PLMI)
  3. improved feature value representation $weight^{SA}$
- Rank cosine value of word pairs via Average Precision (AP).

Dataset	Metric	ADJ		NOUN		VERB	
		SYN	ANT	SYN	ANT	SYN	ANT
ViCon	PPMI	0.70	0.38	0.68	0.39	0.69	0.38
	PLMI	0.59	0.44	0.61	0.42	0.63	0.41
	$weight^{SA}$	**0.93***	**0.31***	**0.94***	**0.31**	**0.96**	**0.31**
	PPMI + SVD	0.76	0.36	0.66	0.40	0.81	0.34
	PLMI + SVD	0.49	0.51	0.55	0.46	0.51	0.49
	$weight^{SA}$ + SVD	**0.91***	**0.32***	**0.81***	**0.34***	**0.92***	**0.32***
LexCon[a]	PLMI	0.56	0.46	0.60	0.42	0.62	0.42
	$weight^{SA}$	**0.75**	**0.36**	**0.66**	**0.40**	**0.71**	**0.38**
	PLMI + SVD	0.55	0.46	0.55	0.46	0.58	0.44
	$weight^{SA}$ + SVD	**0.76***	**0.36***	**0.66**	**0.40**	**0.70***	**0.38***

$\chi^{2,*}$ $p < .001$
[a]Corresponding dataset in English