Word Sense Determination from Wikipedia Data Using Neural Networks

SJSU SAN JOSÉ STATE UNIVERSITY

Advisor
Dr. Chris Pollett
Committee Members
Dr. Jon Pearce
Dr. Suneuy Kim

By
Qiao Liu

Agenda

- Introduction
- Background
- Model Architecture
- Data Sets and Data Preprocessing
- Implementation
- Experiments and Discussions
- Conclusion and Future Work

Introduction

- Word sense disambiguation is the task of identifying which sense of an ambiguous word is used in a sentence.

  *in 1890, he became custodian of the Milwaukee public museum where he collected **plant** specimens for their greenhouse*

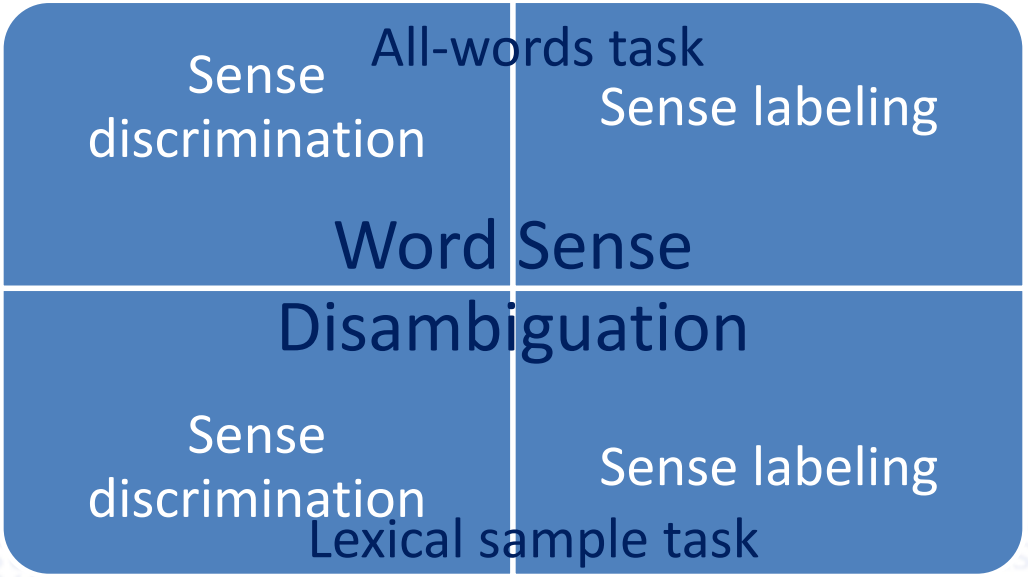
  *…… send collected fluid to a municipal sewage treatment **plant** or a commercial wastewater treatment facility*

- Word sense disambiguation is useful in natural language processing tasks, such as speech synthesis, question answering, and machine translation.

Introduction

Project purpose

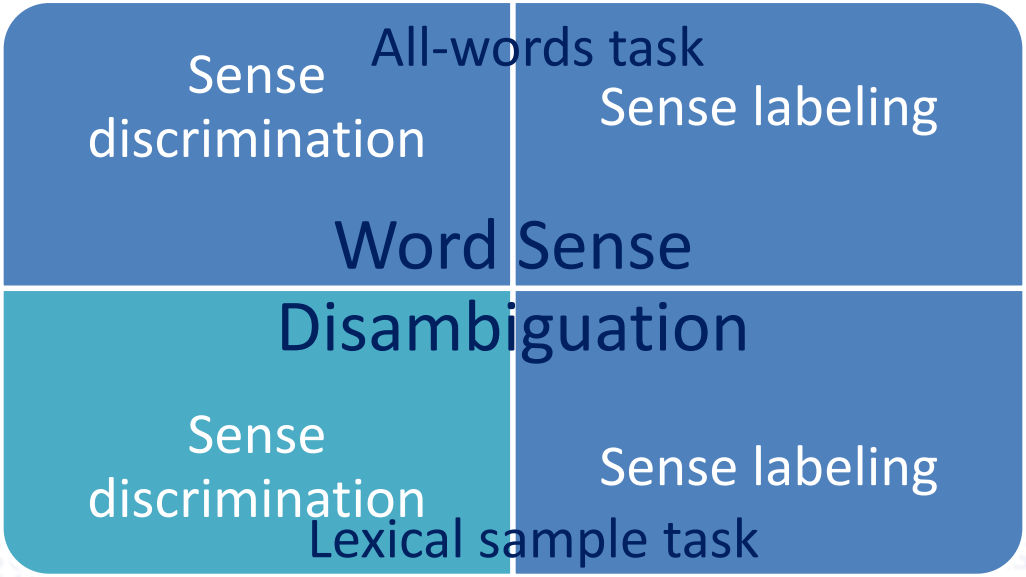
- Two variants of word sense disambiguation task:
  - lexical sample task
  - all-words task
- Two subtasks:
  - sense discrimination
  - sense labeling



Introduction

Project purpose

- Two variants of word sense disambiguation task:
  - **lexical sample task**
  - all-words task
- Two subtasks:
  - **sense discrimination**
  - sense labeling



Existing Work

Background

Approach 1: Dictionary-based

Given a target word t to be disambiguated in Context c.

1. retrieve all the sense definitions for t from a dictionary.
2. select the sense s whose definition have the most overlap with c of t.

- This approach requires a hand-built machine readable semantic sense dictionary.

Approach 2: Supervised machine learning

1. Extract a set of features from the context of the target word.
2. Use the feature to train classifiers that can label ambiguous words in new text.

- This approach requires costly large hand-built resources, because each ambiguous word need be labelled in training data.
- A semi-supervised approach was proposed in 1995 by Yarowsky. In this approach, they do not rely on a large hand-built data, due to using bootstrapping to generate dictionary from a small hand-labeled seed-set.

Approach 3: Unsupervised machine learning

Interpret the sense of the ambiguous word as clusters of similar contexts. Contexts and words are represented by a high-dimensional, real-valued vector using co-occurrence counts.

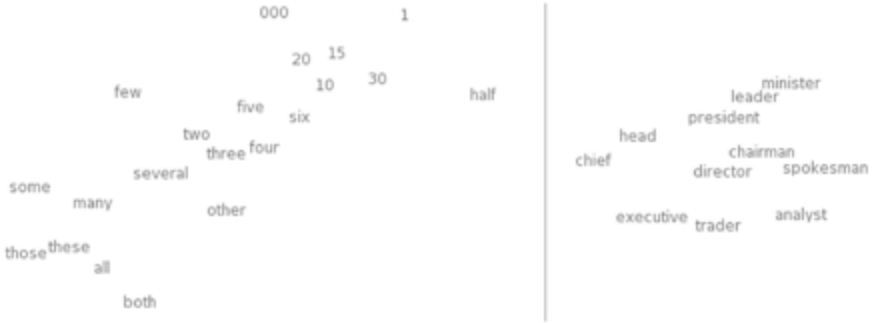
- In our project, we use a modification of this approach:
- Word embeddings are trained using Wikipedia pages.
- Word vectors of contexts computed by these embedding are then clustered.
- Given a new word to disambiguate, we use its context and the word embedding to find a word vector corresponding to this context. Then we determine the cluster it belongs.
- In related work, Schütze used a data set taken from the New York Times News Service and did clustering but with a different kind of word vector.

Background

- Word embeddings

A word embedding is a parameterized function mapping words in some language to high-dimensional vectors (perhaps 200 to 500 dimensions)

word → R^n
W("plant") = [0.3, -0.2, 0.7, …]
W("crane") = [0.5, 0.4 -0.6, …]

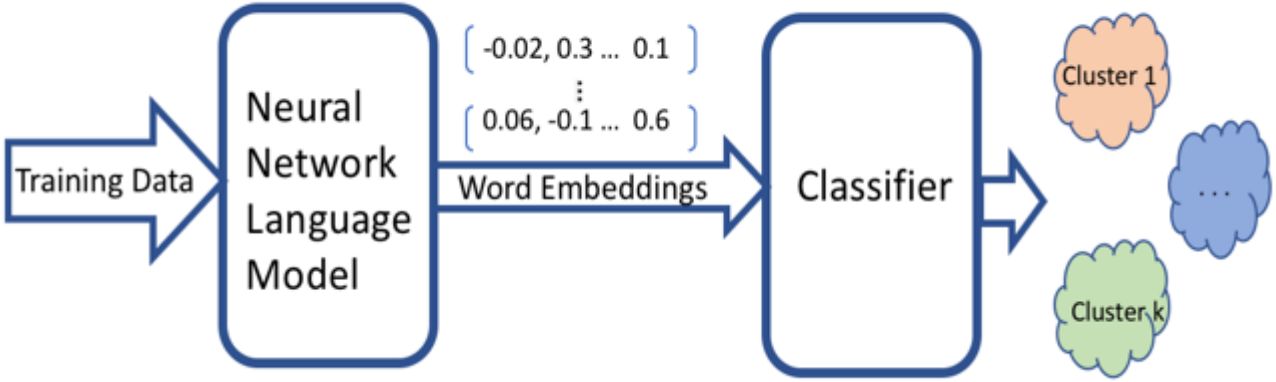


Model Architecture

- Many NLP tasks take the approach of first learning a good word representation on a task and then using that representation for other tasks. We used this approach for the word sense determination task.

Model Architecture

- Learn a good word representation of a task and then using that representation for other tasks.



- We used the Skip-gram model as the neural network language model layer

Model Architecture

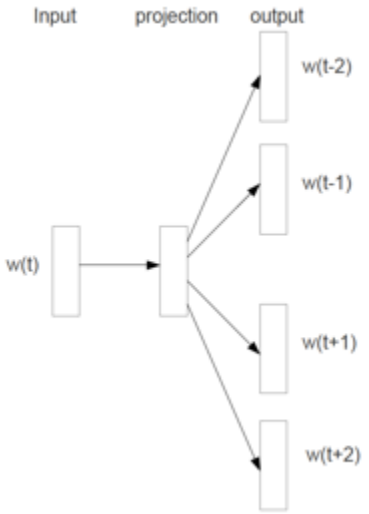
Skip-gram Model Architecture

- The training objective was to learn word embeddings good at predicting the context words in a sentence.
- We trained the neural network by feeding it word pairs of target word and context word found in our training dataset.

$$J'(\theta) = \prod_{t=1}^{V} \prod_{\substack{-m \le j \le m \\ j \neq 0}} p(w_{t+j}|w_t;\theta)$$

$$J(\theta) = -\frac{1}{V} \sum_{t=1}^{V} \sum_{\substack{-m \le j \le m \\ j \neq 0}} \log\left(p(w_{t+j}|w_t;\theta)\right)$$

$$p(w_o|w_t) = \frac{\exp(w_o{}^T w_t)}{\sum_{j=1}^{V} \exp(w_j{}^T w_t)}$$



Model Architecture

- k-means clustering

k-means is a simple unsupervised classification algorithm. The aim of the k-means algorithm is to divide m points in n dimensions into k clusters so that the within-cluster sum of squares is minimize

The distributional hypothesis says that similar words appear in similar contexts [9, 10]. Thus, we can use k-means to divide all vectors of context into k clusters.

SJSU SAN JOSÉ STATE UNIVERSITY

Data Sets and Data Preprocessing

- Data source
  https://dumps.wikimedia.org/enwiki/20170201/
  The `pages-articles.xml` of Wikipedia data dump contains current version of all article pages, templates, and other pages.

- Training data for model
  Word pairs: (target word, context word)

Sentence	Training samples (window size = 2)
natural language processing projects are fun	(natural, language), (natural, processing)
natural **language** processing projects are fun	(language, natural), (language, processing), (language, projects)
natural language **processing** projects are fun	(processing, natural), (processing, language), (processing, projects)
natural language processing **projects** are fun	(projects, language), (projects, processing), (projects, are), (projects, fun)
natural language processing projects **are** fun	(are, processing), (are, project), (are, fun)
natural language processing projects are **fun**	(fun, projects), (fun, are)

SJSU SAN JOSÉ STATE UNIVERSITY

Data Set and Data Preprocessing

Steps to process data:

- Extracted 90M sentences

```
< a line of light emanating from a lighthouse makes one revolution every 10 seconds >
< the lighthouse is located 4km off a straight shoreline >
< how fast does the light move along the shoreline when it forms a 45 degree angle to
a line from the lighthouse perpendicular to the shoreline >
```

- Counted words, created a dictionary and a reversed dictionary

```
UNK:-1              UNK:0
the:102405433       the:1
>:90887424          >:2
<:90887424          <:3
of:49644024         of:4
```

- Regenerated sentences

```
3 8 156 4 587 0 19 8 6111 958 36 1794 399 299 2912 2
3 1 6111 10 155 0 245 8 2143 11381 2
3 225 2649 314 1 587 718 217 1 11381 47 20 1030 8 2926 598 3768 7 8 156 19 1 6111
11304 7 1 11381 2
```

- Created 5B word pairs

Implementation

The optimizer:

- **Gradient descent** finds the minimum of a function by taking steps proportional to the positive of the gradient. In each iteration of gradient descent, we need to calculate all examples.
- Instead of computing the gradient of the whole training set, each iteration of **stochastic gradient descent** only estimates this gradient based on a batch of randomly picked examples.

We used stochastic gradient descent to optimize the vector representation during training.

Implementation

The parameters:

Parameters	Meaning
VOC_SIZE	The vocabulary size.
SKIP_WINDOW	The window size of text words around target word.
NUM_SKIPS	The number of context words, which will be randomly took to generate word pairs.
EMBEDDING_SIZE	The number of parameters in the word embedding. The size of the word vector.
LR	The learning rate of gradient descent
BATCH_SIZE	The size of each batch in stochastic gradient descent. Running one batch is one step.
NUM_STEPS	The number of training step.
NUM_SAMPLE	The number of negative samples.

Implementation

Tools and packages:

- TensorFlow r1.4
- TensorBoard 0.1.6
- Python 2.7.10
- Wikipedia Extractor v2.55
- sklearn.cluster [15]
- numpy

SJSU SAN JOSÉ STATE UNIVERSITY

Experiments and Discussions

The experimental results are compared with Schütze's unsupervised learning approach in 1998:

- Schütze used a data set (435M) taken from the New York Times News Service. We used the data set extracted from Wikipedia pages (12G).
- Schütze used co-occurrence counts to generate vectors, which had large numbers of vector dimension (1,000/2,000).We used the Skip-gram model to learn a distributed word representation with a dimension of 250.
- Schütze applied singular-value decomposition due to large numbers of vector dimension. Taking advantage of a smaller number of dimension, we did not need to perform matrix decomposition.

SJSU SAN JOSÉ STATE UNIVERSITY

Experiments and Discussions

- We experimented the Skip-gram model with different parameters and selected one word embedding for clustering.

- Skip-gram model parameters

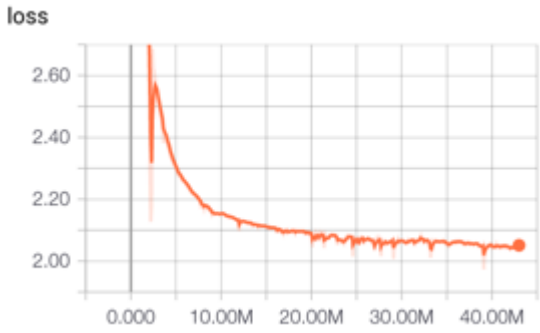
Parameters	VOC_SIZE	SKIP_WINDOW	NUM_SKIPS	EMBEDDING_SIZE
Value	50K	5	10	250
Parameters	LR	BATCH_SIZE	NUM_STEPS	NUM_SAMPLE
Value	0.3	256	Set to 6 epochs	3

Table 6: Parameters for our skip-gram model

Experiments and Discussions

Experiment with skip-gram model

- Used “average loss” to estimate the loss over every 100K batches.
- Visualized some words’ nearest words.




```
Nearest to uk: chart, scottish, netherlands, peaked, sales, irish, canada, charted,
Nearest to islands: island, caribbean, waters, cape, coastal, philippines, bay, provinces,
Nearest to settled: lived, arrived, fled, settle, resided, migrated, settlers, stayed,
Nearest to employed: trained, hired, taught, used, recruited, engaged, appointed, worked,
Nearest to models: designs, engines, systems, components, concepts, model, methods, applications,
Nearest to where: whenever, then, if, what, how, when, which, why,
Nearest to editors: admins, users, pages, admin, edits, editing, user, individuals,
Nearest to island: islands, peninsula, beach, bay, shore, coastal, coast, harbour,
Nearest to style: revival, styles, gothic, classical, elements, distinctive, architecture, genre,
Nearest to moth: beetle, moths, flowering, species, butterfly, beetles, mollusk, bee,
Nearest to leaving: joining, leave, returning, left, entering, losing, seeing, moving,
Nearest to single: double, separate, full-length, solo, disc, cd, promotional, vinyl,
Nearest to labor: labour, socialist, economic, liberal, party, agriculture, democratic, workers,
Nearest to municipal: county, borough, provincial, administrative, city, metropolitan, municipality, regional,
Nearest to convention: congress, treaty, conventions, constitution, delegate, meeting, forum, declaration,
```

Experiments and Discussions

Experiment with classifying word senses

- Clustered the contexts of the occurrences of given ambiguous word into two/three coherent groups.
- Manually assigned labels to the occurrences of ambiguous words in the test corpus, and compare them with machine learned labels to calculate accuracy.
- Before word sense determination, we assigned all occurrences to the most frequent meaning, and used the fraction as the baseline.

human label	label	sentences
living	0	< known locally as UNK is a critically endangered species of flowering plant endemic to the island of mauritius >
living	0	< the fungus provides benefits to the plant which can include increased water or nutrient uptake and protection from UNK insects >
living	0	< a red data book of rare and endangered plant species >
living	1	< and it is from this plant that all the other known specimens in the uk were derived >
factory	0	< the dublin plant formulas use of sugar made it popular among soda fans >
factory	1	< a former printing plant turned into a community arts space in the late 1980s >
factory	1	< the imperium plant has been hit hard by the economic downturn and the drastic changes in the cost of petroleum fuels and biodiesel UNK >

$$\text{accuracy} = \frac{Number\ of\ instances\ with\ correct\ machine\ learned\ sense\ label}{The\ total\ number\ of\ test\ instances}$$

SJSU SAN JOSÉ STATE UNIVERSITY

Experiments and Discussions

- "Schütze's baseline" column gives the fraction of the most frequent sense in his data sets.
- "Schütze's accuracy" column gives the results of his disambiguation experiments with local terms frequency if applicable.
- We got better accuracy out of experiments with "capital" and "plant".
- However, the model cannot determine the senses of word "interest" and "sake", which has a baseline over 85% in our data sets.

word	senses	Training	Test	Schütze's baseline	Schütze's accuracy	Baseline	Accuracy
capital	Stock of goods/ Seat of government	179,793	100	64%	71%	59%	79%
plant	living/factory	164,858	100	54%	64%	59%	76%
ruling	an authoritative decision to exert control / influence			60%	84%	63%	70%
crane	bird/ machine/ person name	6,655	100	-	-	46%	68%
interest	A feeling of special attention/ A charge for borrowed money	112,903	100	58%	90%	86%	49%
train	benefit/drink	9,290	100	74%	69%	91%	57%

Experiments and Discussions

Discussions

- Our data sets (12G) are much larger than Schütze's data sets (435M). For example, the size of his training set for word "capital" is 13,015, and ours is 179,793. The larger data sets might have helped to increase the accuracy for some words.
- We also observed that when the baseline is high (>= 85%), the model cannot determine the senses of the word. The performance of unsupervised learning relies on sufficient information from the training data. However, the model didn't get trained with sufficient data carrying less frequent meanings.
- The size of the training data, and the distribution of the senses of the target word has significant influent to the performance of the model.

Conclusion and Future Work

Conclusion

- In this project, we utilized the distributional word representation and the distributional hypothesis to build a modular model to classify the senses of ambiguous words.

- Our experiments showed our model performed well when an ambiguous word had each sense accounts for than 20% of occurrences in the training data set.

Conclusion and Future Work

Future Work

- Optimize the classifier. One possible approach might be using weighted sum of contexts by taking IDF into account.
- Extend and experiment this approach to other models with different classifiers. The classifier which works well when occurrences are skewed to one class might improve the accuracy for words with large portion of occurrences are using the most frequent sense.
- Tokenize the corpus, we could reduce the time cost of training by reducing vocabulary size.

SJSU SAN JOSÉ STATE UNIVERSITY

References

- Y. Bengio, R. Ducharme, P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 2003.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.
- G.E. Hinton, J.L. McClelland, D.E. Rumelhart. Distributed representations. In: Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations, MIT Press, 1986.
- T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, 2007.
- David E Rumelhart, Geoffrey E Hintont, and Ronald J Williams. Learning representations by backpropagating errors. Nature, 323(6088):533–536, 1986.
- H. Schwenk. Continuous space language models. Computer Speech and Language, vol. 21, 2007.
- T. Mikolov, A. Deoras, S. Kombrink, L. Burget, J. Cˇ ernocky´. Empirical Evaluation and Combination of Advanced Language Modeling Techniques, In: Proceedings of Interspeech, 2011.

SJSU SAN JOSÉ STATE UNIVERSITY

References

- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, 2013a.
- James R. Curran and Marc Moens. Improvements in automatic thesaurus extraction. In Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition, pages 59–66. 2002.
- Patrick Pantel and Dekang Lin. Discovering word senses from text. In Proc. Of SIGKDD-02, pages 613–619, New York, NY, USA. ACM. 2002.
- Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of SIGDOC, pages 24-26, 1986.
- Olah, Christopher. Deep Learning, NLP, and Representations. Retrieved from http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/. 2014
- Hartigan, J. A. and Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics). 28 (1): pages 100–108, 1979.
- Schütze, Hinrich. Dimensions of meaning. In Proceedings of Supercomputing'92, pages 787-796, 1992.

SJSU SAN JOSÉ STATE UNIVERSITY

References

- Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.
- Michael U Gutmann and Aapo Hyv¨arinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. The Journal ofMachine Learning Research, 13:307–361, 2012.
- Bottou L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent. In: Lechevallier Y., Saporta G. (eds) Proceedings of COMPSTAT'2010. Physica-Verlag HD
- TensorFlow Tutorial, tf.nn.nce_loss. Retriveved from https://www.tensorflow.org/api_docs/python/tf/nn/nce_loss. 2017
- McCormick, C, Word2Vec Tutorial Part 2 - Negative Sampling. Retrieved from http://www.mccormickml.com, 2017, January 11.
- D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods, Proc. 33rd Annual meeting of the ACL, Cambridge, MA, USA, pp 189-196, 1995.
- Schütze, Hinrich, Automatic word sense discrimination, Computational Linguistics, v.24 n.1, March 1998

Questions

SAN JOSÉ STATE UNIVERSITY *powering* SILICON VALLEY

Thank You!

Appendix: Model Architecture

Skip-gram model architecture

- We trained the neural network by feeding it word pairs of target word and context word found in our training dataset.