Proposal Machine Learning And Natural Language Processing On Hotel Reviews

Name: Yi ZHOU

Student Number: 422991

Software development, School of Information

Abstraction:

Keyword:

Contents

[1. Introduction to machine learning 1](#_Toc496525409)

[2. Introduction to natural language processing 1](#_Toc496525410)

[3. Introduction to sentiment analysis 1](#_Toc496525411)

1. Introduction to machine learning

Machine Learning is subfield of Artificial intelligence.

Machine Learning can be classified into 3 categories, supervised machine learning, semi-supervised machine learning, unsupervised machine learning, reinforcement learning.

* 1. Introduction to supervised machine learning
  2. Introduction to semi-supervised machine learning
  3. Introduction to unsupervised machine learning

Unsupervised machine learning contains a few approaches or algorithms. In clustering, k-means, mixture models and hierarchical clustering are all classic approaches to cluster data.

In natural language process area, unsupervised machine learning can be used to learn to morphology. For instance, given set of words, one can bootstrap the process of learning suffixes, stem. Researchers used minimum description length for unsupervised morphology learning. In the area of sentence segmentation, which is the process of dividing written text into meaningful units, such as words, phrases, Unsupervised machine learning is valuable to identify sentence boundary. NLTK, which is a popular third-party library in NLP field, its sentence detection tool uses unsupervised machine learning. In Bootstrapping aggregating, Expectation Maximization algorithm is used with classification algorithm to bootstrap classification, we can learn more details of this method from 2. In disambiguation problem, unsupervised machine learning and semi-supervised learning are used for different disambiguation tasks like word sense disambiguation, entity disambiguation etc. In Machine Translation, Expectation Maximum is applied widely as well.

* + 1. The advantage of unsupervised machine learning

The potential of unsupervised learning is to avoid hand-engineering features and to make use of the massive amounts of easily accessible unlabeled text to learn feature embeddings which are useful for many tasks.

* 1. Introduction to reinforcement learning

Reinforcement learning is the problem faced by an agent that must learn behavior through trial-and-error interactions with a dynamic environment.

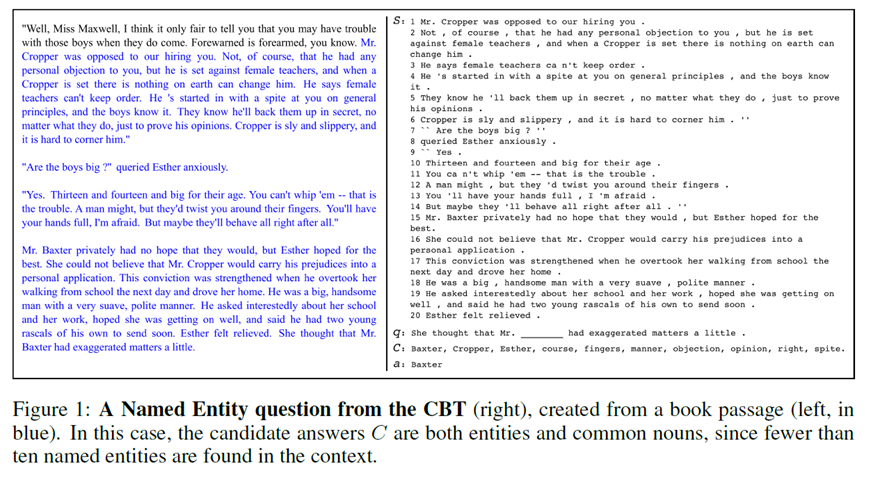
1. Introduction to natural language processing

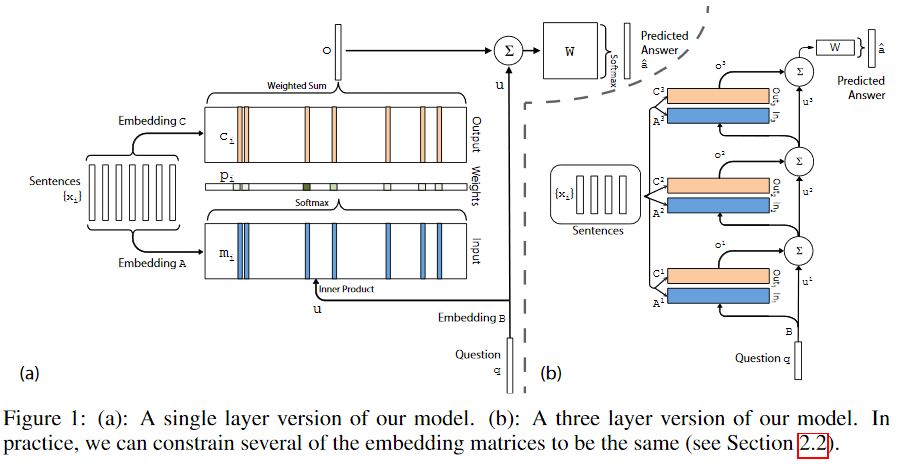
Natural language processing contains different field, including Machine Translate, Question Answering, Opinion Mining, speech recognition.

* 1. Machine translation introduction
  2. Question answering introduction

THE GOLDILOCKS PRINCIPLE: READING CHILDREN’S BOOKS WITH EXPLICIT MEMORY REPRESENTATIONS,

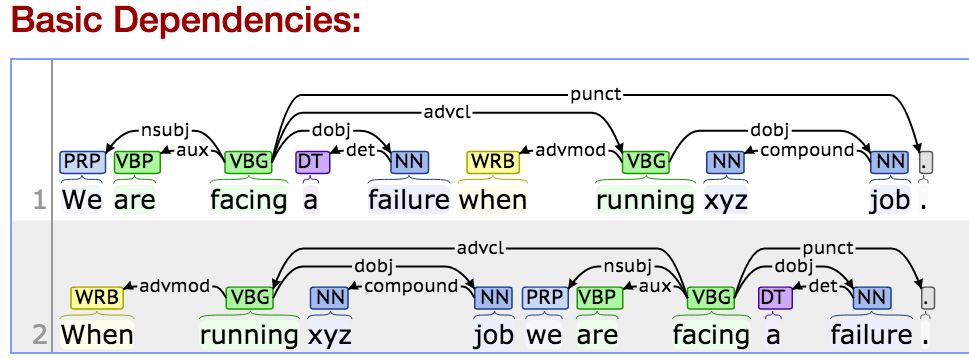
* Randomly selecting 10 nouns or Named entities from paragraph.
* The first 20 sentences are considered as document.
* There 10 answer candidates are in same category.
  + Named Entity
  + Nouns
  + Verbs
  + Preposition





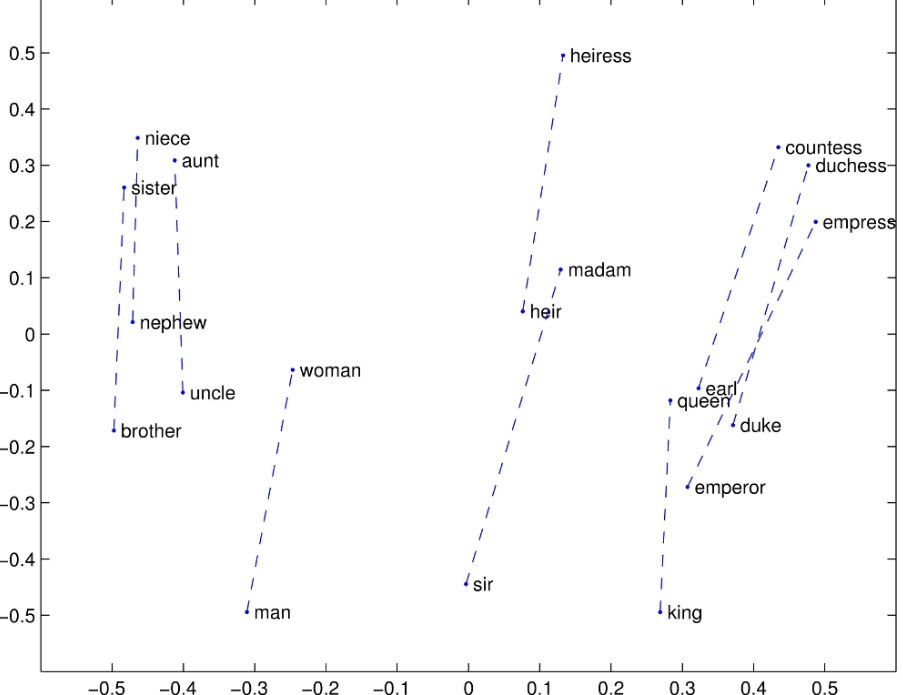
* 1. Speech recognition
  2. Name entity recognition
  3. Dependency parsing

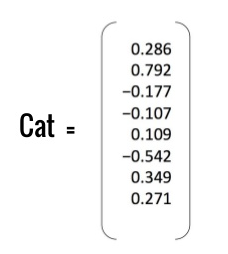
Dependency Parsing – NLP，



* 1. Word2vec introduction

Learning word vectors for sentiment analysis.





1. Introduction to sentiment analysis (opinion mining)

Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media.

The following are the example of sentiment analysis

“The strange thing is that it works”

“Shattered image isn’t complex, it’s just stupid and boring”

“Far from bewitching, the crucible tests the patience”

* 1. Objectives

The Theoreticians

Introduce and review NLP techniques for sentiment analysis

Discuss possible ways to improve challenging task

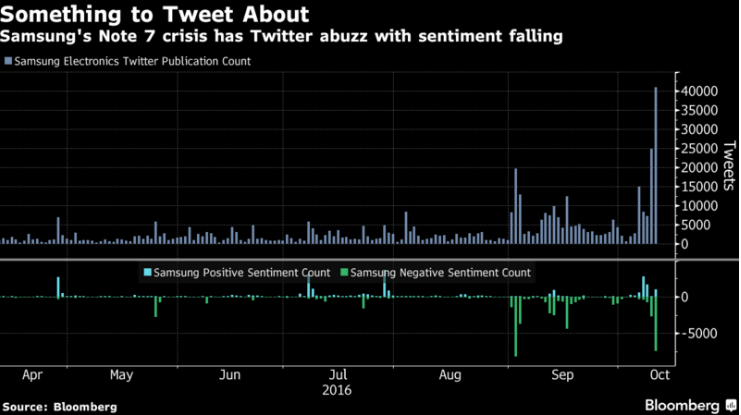
The Practicians

Use NLP for improving/ providing better solutions

Understanding scalability of API, like Google Cloud NLP API

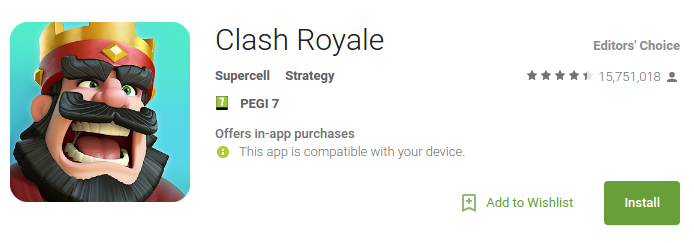
Develop a general purpose classifier with Tools

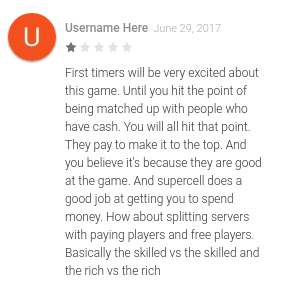
* 1. Twitter sentiment example

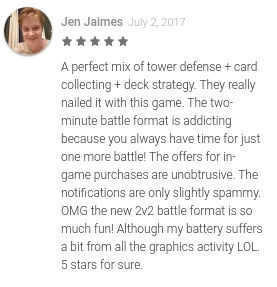
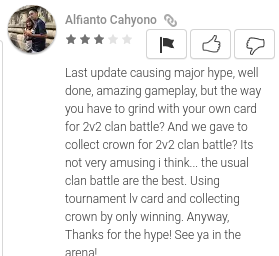


Source link: <https://www.bloomberg.com/news/articles/2016-10-12/samsung-tweets-soared-sentiment-soured-amid-note-7-halt-chart>

* 1. Analyzing App Popularity - Use case







<https://play.google.com/store/apps/details?id=com.supercell.clashroyale&hl=en>

* 1. Potential Use Cases
* Prediction
  + Analyze market trends from Tweets/ Weibo or other news articles
  + Estimating movie success from reviews, blogs etc
* Products
  + How do users like iOS update ? How do players think of apps updating ?
* Politics
  + What do voters think about the a newly proposed policy ?
  1. Why Is Sentiment Analysis Hard ?

Language usage, sarcasm, lingo

“This movie is just brilliant. Actors are first grade but the camera work is shoddy, at best”

“This film runs for a slick 140 mins, I was blown away by how true Einstein was after all”

“Ok lar i dbl check wif da hair dresser already he said wun cut v short. He said will cut until i look nice.”

1. Reference
2. Zhu, X., 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, *2*(3), p.4.
3. Kiss, T. and Strunk, J., 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, *32*(4), pp.485-525.
4. Klein, D. and Manning, C.D., 2005. *The unsupervised learning of natural language structure*. Stanford, CA: Stanford University.
5. Kaelbling, L.P., Littman, M.L. and Moore, A.W., 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, pp.237-285.
6. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P., 2011. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), pp.2493-2537.Vancouver
7. Demiroz, G., Yanikoglu, B., Tapucu, D. and Saygin, Y., 2012, December. Learning domain-specific polarity lexicons. In *Data mining workshops (ICDMW), 2012 IEEE 12th international conference on* (pp. 674-679). IEEE.
8. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. and Dyer, C., 2007, June. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 177-180). Association for Computational Linguistics.
9. Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.
10. Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), pp.257-286.
11. Tan, S. and Zhang, J., 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, *34*(4), pp.2622-2629.
12. Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer US.
13. Ding, X., Liu, B. and Zhang, L., 2009, June. Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1125-1134). ACM.
14. Wang, S., Chen, Z. and Liu, B., 2016, April. Mining aspect-specific opinion using a holistic lifelong topic model. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 167-176). International World Wide Web Conferences Steering Committee.
15. Hill, F., Bordes, A., Chopra, S. and Weston, J., 2015. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv preprint arXiv:1511.02301*.
16. Ye, Q., Zhang, Z. and Law, R., 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, *36*(3), pp.6527-6535.
17. Kang, H., Yoo, S.J. and Han, D., 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, *39*(5), pp.6000-6010.
18. Ye, Q., Zhang, Z. and Law, R., 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, *36*(3), pp.6527-6535.
19. Olsson, F., 2009. A literature survey of active machine learning in the context of natural language processing.
20. Liu, B., 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), pp.1-167.
21. Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
22. Pennington, J., Socher, R. and Manning, C., 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
23. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, *521*(7553), pp.436-444.
24. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P. and Bengio, S., 2010. Why does unsupervised pre-training help deep learning?. *Journal of Machine Learning Research*, *11*(Feb), pp.625-660.
25. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C., 2011, June. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 142-150). Association for Computational Linguistics.
26. Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
27. Goldberg, Y., 2016. A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Intell. Res.(JAIR)*, *57*, pp.345-420.
28. Pang, B. and Lee, L., 2005, June. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115-124). Association for Computational Linguistics.
29. Ghorpade, T. and Ragha, L., 2012, October. Featured based sentiment classification for hotel reviews using NLP and Bayesian classification. In *Communication, Information & Computing Technology (ICCICT), 2012 International Conference on* (pp. 1-5). IEEE.