

Title	語クラスターとランキングモデルを用いる 情報更新タスクの扱いに関する研究
Author(s)	PHAM, QUANG NHAT MINH
Citation	
Issue Date	2010-03
Type	Thesis or Dissertation
Text version	author
URL	<a href="http://hdl.handle.net/10119/8932">http://hdl.handle.net/10119/8932</a>
Rights	
Description	Supervisor:Professor Akira Shimazu, 情報科学研究科, 修士

# Treating Information Update Tasks with Word Clusters and Ranking Models

Pham Quang Nhat Minh (0810054)

School of Information Science,  
Japan Advanced Institute of Science and Technology

February 09, 2010

**Keywords:** Information Update, IR, Word Clustering, Ranking Models.

The task of updating information is a significant task in the context that many applications require documents to be updated quite often. In legal domain, it is an important task because of the massive number of legal updates and the cross-reference problem. Our research copes with a special case of the information update task, the information insertion task which aims to determine the most appropriate location to insert a piece of new information into an existing document.

In [6], the information insertion task was formulated as a hierarchical ranking problem. Each document is represented as a hierarchy of sections, paragraphs. Then, the insertion is operated over that hierarchical tree. To determine the best paragraph in the document to add a new sentence, all paragraphs of the document are ranked by a ranking function computed for each insertion sentence/paragraph pair and then, the paragraph with the highest score will be chosen. The ranking function for each insertion sentence/paragraph pair is computed based on a weight vector learned from training data. The training procedure was implemented in an online learning framework with the Perceptron algorithm [13, 8].

We investigated ranking models for the information insertion task on two datasets: Wikipedia insertion dataset obtained from [6] and Legal dataset built by ourselves. The Legal dataset was built from the United States Code which is a compilation and codification of general and permanent

federal law of the United States. The experiment results show that when the deep semantics analysis for texts is not performed, the ranking models with the supervised approach outperform the unsupervised methods for the information insertion task.

In Natural Language Processing, semantic relations between words can be exploited when measuring semantic text similarity of two text segments. In our research, we proposed a method of measuring topical overlap between two text segments, which incorporates word clusters [5, 21, 24], and used these similarity measures as additional semantic features in the learning model. In our method, first, word clusters are derived from unlabeled data. Then, extracted word clusters are used as intermediate representations of words to exploit the semantic similarity and semantic relatedness between words which are different in surface forms but semantically related. The semantic text similarity scores are computed with various kinds of similarity functions. Our results show that combining cluster-based features with baseline features can boost the performance of the information insertion task on two datasets. In the best setting, we obtained 40.4% accuracy of choosing paragraphs on Wikipedia dataset and 52.3% accuracy of choosing section on Legal dataset.

## References

- [1] A. Baeza-Yates, R., Ribeiro-Neto B. (1999). *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA.
- [2] Baker, L. D., McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 96–103.
- [3] Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In: *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.
- [4] Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y. (2003). Distributional word clusters vs. words for text categorization. *The Journal of Machine Learning Research*.
- [5] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), pp. 467-479.
- [6] Chen, E., Snyder, B., and Barzilay, R. (2007). Incremental text structuring with online hierarchical ranking. In *Proceedings of the EMNLP*, pp. 83–91.
- [7] Chen, E. (2008). *Discourse Models for Collaboratively Edited Corpora*. Masters thesis, Massachusetts Institute of Technology.
- [8] Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the EMNLP*, pp. 1-8.
- [9] Corley, C., Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor.

- [10] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y. (2006). Online Passive-Aggressive Algorithms. *The Journal of Machine Learning Research*, vol. 7, pp. 551–585.
- [11] Dagan, I., Glickman, O., Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In: Quionero-Candela, J., Dagan, I., Magnini, B., dAlch-Buc, F. (eds.) *MLCW 2005. LNCS (LNAI)*, vol. 3944, pp. 177-190. Springer, Heidelberg.
- [12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, vol. 41, pp. 391–407.
- [13] Freund, Y., and Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277-296.
- [14] Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B. (2007). The Third PASCAL Recognizing Textual Entailment Challenge. In: *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.
- [15] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL98 Workshop on Usage of WordNet for NLP*.
- [16] Jurafsky, D., and Martin, J. H. (2008). *Speech and Language Processing*. Prentice-Hall, New Jersey, USA.
- [17] Katayama, T. (2007). Legal engineering an engineering approach to laws in e-society age. In: *Proc. of the 1st Intl. Workshop on JURISIN*.
- [18] Katayama, T., Shimazu, A., Tojo, S., Futatsugi, K., Ochimizu, K. (2008). e-Society and legal engineering (in Japanese). *Journal of the Japanese Society for Artificial Intelligence* 23(4), 529-536.
- [19] Kim, H. D., and Zhai, C. (2009). Generating Comparative Summaries of Contradictory Opinions in Text. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM'09)*, Hongkong, pp. 385–394.

- [20] Kimura, Y., Nakamura, M., Shimazu, A. (2008). Treatment of legal sentences including itemized and referential expressions towards translation into logical forms. In: Proc. of the 2nd Intl. Workshop on JURISIN, pp. 73-82.
- [21] Koo, T., Carreras, X., and Collins, M. (2008). Simple semi-supervised dependency parsing. In Proceedings of ACL-08, pp. 595-603.
- [22] Lapata, M., and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In IJCAI, pages 1085-1090.
- [23] Li, W., and McCallum, A. (2005). Semi-supervised sequence modeling with syntactic topic models. In Proceedings of Twentieth National Conference on Artificial Intelligence, pp. 813–818.
- [24] Liang, P. (2005). Semi-Supervised Learning for Natural Language. Masters thesis, Massachusetts Institute of Technology.
- [25] Mihalcea, R., Moldovan, D. (2000). Semantic indexing using WordNet senses. In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, October 08–08, 2000, Hong Kong.
- [26] Miller, George A. WordNet - About Us. (1009). WordNet-Princeton University. Website: <http://wordnet.princeton.edu>
- [27] Miller, S., Guinness, J., and Zamanian, A. (2008). Name tagging with word clusters and discriminative training. In Proceedings of HLT-NAACL, pp. 337-342.
- [28] Nakamura, M., Nobuoka, S., Shimazu, A. (2008). Towards translation of legal sentences into logical forms. In Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T., eds.: New Frontiers in Artificial Intelligence: JSAI 2007 Conference and Workshops, Miyazaki, Japan, June 18-22, 2007, Revised Selected Papers. Volume 4914 of Lecture Notes in Computer Science., Springer, pp. 349-362.
- [29] Nakamura, M., Kimura, Y., Pham, M. Q. N., Nguyen, M. L., and Shimazu, A. (2008). Treatment of Legal Sentences Including Itemization

Written in Japanese, English and Vietnamese. In Proc. of the EMALP Workshop, PRICAI 2008, Hanoi, Vietnam, pp.102–113.

- [30] Ogawa, Y., Inagaki, S., Toyama, K. (2008). Automatic Consolidation of Japanese Statutes Based on Formalization of Amendment Sentences. In: Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T. (eds.) JSAI 2007. LNCS, vol. 4914, pp. 349-362. Springer, Heidelberg.
- [31] Pham, M. Q. N., Nguyen, M. L., Shimazu, A. (2009). Incremental Text Structuring with Word Clusters. In Proceedings of the Conference of the Pacific Association for Computational Linguistics 2009, Hokkaido, Japan, pp. 109–114.
- [32] Pham, M. Q. N., Nguyen, M. L., Shimazu, A. (2010). The Information Insertion Task with Intermediate Word Representation. The 16<sup>th</sup> NLP Annual Meeting, Tokyo, 2010, March. (to appear)
- [33] Ponte J. M., Croft W. B. (1998). A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275–281.
- [34] Qiu, Y., Frei, H. (1993). Concept based query expansion. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 160–169.
- [35] Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACL Workshop on Summarization Seattle, WA.
- [36] Robertson, S., Zaragoza, H., Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In: Proc. of the thirteenth ACM international conference on Information and knowledge management, pp. 42–49.
- [37] Toutanova, K., and Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical

Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63–70.

- [38] Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252–259.
- [39] Preliminary Recommendations on Semantic Encoding Interim Report (1998). Retrieved January 2010 from the Website of Expert Advisory Group on Language Engineering Standards: <http://www.ilc.cnr.it/EAGLES96/rep2/node37.html>
- [40] e-Government. Retrieved from the Wikipedia: [http://en.wikipedia.org/wiki/E\\_government](http://en.wikipedia.org/wiki/E_government)
- [41] United States Code. Retrieved from the Website of the U.S. Government Printing Office: <http://www.gpoaccess.gov/uscode/about.html>
- [42] Website of Office of the Law Revision Counsel. The United States Code. Retrieved from <http://uscode.house.gov/lawrevisioncounsel.shtml>.
- [43] Website of Essem Educational. Text Coherence and Cohesion. Retrieved from <http://www.readability.biz/Coherence2.html>