# Research Proposal

*Named-entity recognition and relation extraction in unstructured data using neural models*

**Kurt Junshean Espinosa**
Department of Computer Science
University of the Philippines Cebu
kpespinosa@up.edu.ph

August 16, 2015

# Contents


**1 Overview of the research** **1**

**2 Positioning of the research** **1**

**3 Research design methodology** **2**

# 1 Overview of the research

As human activities affect biodiversity, in the same way, biodiversity impacts human life. Therefore, the better we understand biodiversity, the sooner we can lessen the negative impact or increase the positive impact to human life. However, as pointed out by Subramaniam, et. al.[1], because of the large amounts of data being generated day by day, it is almost impossible to keep track of all the information and present them in a way useful to researchers and decision-makers, thus, there is a need for an automated information extraction for timely dissemination of information.

This problem has been observed by Beaman, et.al. in their study[2] that one of the least tapped sources of biodiversity knowledge is the collection locations, dates, species identification and other information on over a billion natural history specimen labels worldwide and only a very small fraction of these have been digitized and the information added to databases. Clearly, there is a huge gap that needs to be addressed.

The aim of this study is to bridge this gap of having so much data on biodiversity available but as it is now has only been scarcely useful. To achieve this goal, novel research on the extraction and normalisation of entities will be done for biodiversity at large scale. Moreover, improvement on current methods of relation extraction will be investigated.

This study will therefore contribute to the development of a semantic search system to help researchers and the public study scientific documents on biodiversity as explained in [3]. In particular, this will be helpful in formulation of environmental policies, and the discovery of new natural products that can potentially provide medicinal benefits or this may help advance the cancer research[4]. In general, this study hopes to advance question-answering and text-mining in text.

# 2 Positioning of the research

The interdisciplinary nature of this research requires the close communication among those involved. For example, domain experts in biodiversity, social sciences, linguistics, and computer science have to be able to work together while each of them contributing their expertise. Particularly, vocabularies for biodiversity have to built by domain experts because they will be needed during text mining.

Grishman[5] has provided a thorough discussion on the capabilities and challenges of Information Extraction(IE). In the field of biomedical text mining, Cohen and Hersh made a survey[6] of the current work on named entity recognition, text classification, terminology extraction, relationship extraction and hypothesis generation. Furthermore, Ananiadou[7] also discussed the techniques and tools used for doing text mining in biomedicine.

In text mining, among the fundamental tasks are named entity recognition (NER) and relation extraction(RE). A survey of named-entity recognition (NER) techniques used are discussed by Nadeau, et. al[8] and Sharnagat[9] and in the biomedical domain by Sondhi[10]. Research also on named-entity extraction(NEE) from unstructured text applying semantic parser and coreference solved was done by Exner and Nugues[11]. Many researches also have experimented with using knowledge base such as wikipedia as training data[12], comparing with other corpora[13], as an external knowledge[14], and for NER in social media[15]. Unsupervised approaches for NEE has been applied to the web[16] and semi-supervised learning method to biomedical text mining[17]. Disambiguation problems have also been investigated using wikipedia[18], [19]. NER among others have been applied to the medical domain for recognizing drug[20].

The task of relation extraction, on the other hand, is to predict semantic relations between pairs of entities. State-of-the-art of event or relation extraction is discussed in [21], [22]. The most representative methods for relation classification use supervised paradigm[23], [24], [25], [26]. Supervised methods are grouped into feature-based and kernel-based methods. In feature-based methods, sequences and parse trees are investigated as clues and are converted into feature vectors[27], [28] but finding the suitable feature set is still a problem. A survey of kernel-based methods is discussed by Moncecchi, et.al.[29]. Generally, kernel-based methods still suffers from lack of sufficient labeled data for training. Bootstrapping based approaches, however, result in the discovery of large number of patterns and relations[22]. Researches on relation extraction rest on the distributional hypothesis theory[30] which indicates that words that occur in the same context tend to have similar meanings. In consequence, it is assumed that the pairs of entities that occur in similar contexts tend to have similar relations. Semi-supervised methods have also been investigated such as Snowball[31], KnowItAll[16], TextRunner[32] in view of the limited labeled data.

## 3 Research design methodology

Given the unique nature of the biodiversity domain, evaluation of the methods will be done closely with domain experts. Particularly, it will be explored in this study how to take advantage of the huge dataset in developing features and representations automatically. Current advances in neural methods such as deep learning such as word embeddings[33], [34], [35], [36], [37] will be investigated .

It will have to be studied also which neural architectures would perform best on NEE and RE tasks. There have been studies on learning representations with recursive neural networks[38] and relation classification using convolutional deep neural networks[39] and improving them by connecting them with knowledge bases[40]. Word representations can also be employed for domain adaptation of relation extraction[41]. Neural network based models have also shown promise for modelling reading comprehension[42].

## References

[1] L. V. Subramaniam, S. Mukherjea, P. Kankar, B. Srivastava, V. S. Batra, P. V. Kamesam, and R. Kothari, “Information extraction from biomedical literature: methodology, evaluation and

an application," in *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 410–417, ACM, 2003.

[2] R. S. Beaman, N. Cellinese, P. B. Heidorn, Y. Guo, A. M. Green, and B. Thiers, "Herbis: Integrating digital imaging and label data capture for herbaria," *Botany 2006, Chico, CA*, 2006.

[3] "Transatlantic digging into data challenge 2013 winners announced." http://miningbiodiversity.org/project-details/, 2014. [Online; accessed 15-August-2015].

[4] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, and B. Shen, "Biomedical text mining and its applications in cancer research," *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 200 – 211, 2013.

[5] R. Grishman, "Information extraction:capabilities and challenges." http://cs.nyu.edu/grishman/tarragona.pdf, 2012. [Online; accessed 16-August-2015].

[6] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[7] S. Ananiadou and J. McNaught, *Text mining for biology and biomedicine*. Citeseer, 2006.

[8] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[9] R. Sharnagat, "Named entity recognition:a literature survey." http://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf, 2014. [Online; accessed 16-August-2015].

[10] P. Sondhi, "A survey on named entity extraction in the biomedical domain." http://sifaka.cs.uiuc.edu/~sondhi1/survey1.pdf. [Online; accessed 16-August-2015].

[11] P. Exner and P. Nugues, "Entity extraction: From unstructured text to dbpedia rdf triples," in *The Web of Linked Entities Workshop (WoLE 2012)*, 2012.

[12] J. Nothman, J. R. Curran, and T. Murphy, "Transforming wikipedia into named entity training data," in *Proceedings of the Australian Language Technology Workshop*, pp. 124–132, 2008.

[13] D. Balasuriya, N. Ringland, J. Nothman, T. Murphy, and J. R. Curran, "Named entity recognition in wikipedia," in *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 10–18, Association for Computational Linguistics, 2009.

[14] J. Kazama and K. Torisawa, "Exploiting wikipedia as external knowledge for named entity recognition," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 698–707, 2007.

[15] A. Gattani, D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan, "Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach," *Proceedings of the VLDB Endowment*, vol. 6, no. 11, pp. 1126–1137, 2013.

[16] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the web: An experimental study," *Artificial Intelligence*, vol. 165, no. 1, pp. 91 – 134, 2005.

[17] T. Munkhdalai, M. Li, K. Batsuren, H. A. Park, N. H. Choi, and K. H. Ryu, "Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations," *Journal of cheminformatics*, vol. 7, no. Suppl 1, p. S9, 2015.

[18] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data.," in *EMNLP-CoNLL*, vol. 7, pp. 708–716, 2007.

[19] R. C. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation.," in *EACL*, vol. 6, pp. 9–16, 2006.

[20] I. Korkontzelos, D. Piliouras, A. W. Dowsey, and S. Ananiadou, "Boosting drug named entity recognition using an aggregate classifier," *Artificial Intelligence in Medicine*, pp. –, 2015.

[21] F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong, "An overview of event extraction from text," in *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011)*, vol. 779, pp. 48–57, Citeseer, 2011.

[22] N. Bach and S. Badaskar, "A review of relation extraction," *Literature review for Language and Statistics II*, 2007.

[23] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *The Journal of Machine Learning Research*, vol. 3, pp. 1083–1106, 2003.

[24] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 724–731, Association for Computational Linguistics, 2005.

[25] G. ZHOU12, M. Zhang, D. H. Ji, and Q. Zhu, "Tree kernel-based relation extraction with context-sensitive structured parse tree information," *EMNLP-CoNLL 2007*, p. 728, 2007.

[26] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011, Association for Computational Linguistics, 2009.

[27] N. Kambhatla, “Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations,” in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p. 22, Association for Computational Linguistics, 2004.

[28] F. M. Suchanek, G. Ifrim, and G. Weikum, “Leila: Learning to extract information by linguistic analysis,” in *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pp. 18–25, 2006.

[29] G. Moncecchi, J.-L. Minel, and D. Wonsever, “A survey of kernel methods for relation extraction,” in *Workshop on NLP and Web-based technologies*, 2010.

[30] Z. S. Harris, “Distributional structure.,” *Word*, 1954.

[31] E. Agichtein and L. Gravano, “Snowball: Extracting relations from large plain-text collections,” in *Proceedings of the fifth ACM conference on Digital libraries*, pp. 85–94, ACM, 2000.

[32] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland, “Textrunner: open information extraction on the web,” in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 25–26, Association for Computational Linguistics, 2007.

[33] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model,” *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[34] A. Mnih and G. Hinton, “Three new graphical models for statistical language modelling,” in *Proceedings of the 24th international conference on Machine learning*, pp. 641–648, ACM, 2007.

[35] G. E. Hinton, “Learning multiple layers of representation,” *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[36] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008.

[37] J. Turian, L. Ratinov, and Y. Bengio, “Word representations: a simple and general method for semi-supervised learning,” in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384–394, Association for Computational Linguistics, 2010.

[38] R. Socher, C. D. Manning, and A. Y. Ng, “Learning continuous phrase representations and syntactic parsing with recursive neural networks,” in *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pp. 1–9, 2010.

[39] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, “Relation classification via convolutional deep neural network,” in *Proceedings of COLING*, pp. 2335–2344, 2014.

[40] J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier, “Connecting language and knowledge bases with embedding models for relation extraction,” *arXiv preprint arXiv:1307.7973*, 2013.

[41] T. H. Nguyen and R. Grishman, "Employing word representations and regularization for domain adaptation of relation extraction," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 68–74, 2014.

[42] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *CoRR*, vol. abs/1506.03340, 2015.