Mario J. Lorenzo
mario@mjlorenzo.com

Critique of:

Lamurias, A., Sousa, D., Clarke, L. A., & Couto, F. M. (2019). BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC bioinformatics*, *20*(1), 10.

1. Problem Addressed:

Relation extraction and classification is a common Information Extraction task involving the linking of entities that participate in a semantic or syntactic relationship within a span of text. The process of relationship extraction is typically dependent on other NLP tasks such as entity detection, part-of-speech (POS) tagging, or syntactic parsing. There has been significant work resulting in numerous methods that address this problem. These methods can be categorized into knowledge-based, supervised, and semi-supervised (Etzioni, 2008).

The critiqued paper by Lamurias et al (2019) proposes an extension to previous work by B. Xu et al (2018) and Y. Zhang et al (2018) through the introduction of an ontology embedding layer added to a baseline model previously proposed by (Y. Zhang et al, 2018) which produces dense vector representations using a sequence of ontological ancestors for concepts resolved from entities within each input span. This approach leverages an RNN/LSTM featuring a multi-channel stacked neural network architecture that combines multiple embedding sources including the proposed ontological concept embedding method (further elaborated within the Methodology section of this critique).

The hypothesis for this work is supported by the success of current state-of-the-art relation extraction methods that leverage LSTM along with Word embeddings layers (Mikolov, 2013; Pennington, 2014). Lamurias et al postulate that an ontology (or concept) embedding layer can provide additional domain-specific information that is otherwise hidden in the training data. This insight stems from both the well-known success and limitations of word embeddings (discussed further in 'Prior Research and Significance' below).

2. Prior Research and Significance:

The need for improved relation extraction methods is a crucial prerequisite to natural language understanding and an integral aspect of knowledge discovery, Q&A, semantic search engines, and decision support systems (Bach and Badaskar, 2007). This need stems from the challenge presented by the rapidly growing and vast amount of unstructured text available on the web in the form of journals, blogs, news articles, and government documents.

Early relationship extraction methods were based on pattern-matching (Riloff and Jones, 1999), lexico-syntactic patterns (Hearst et al, 1992), cascading grammars (Boguraev, 2004), relational (Reiss, 2008; Krishnamurthy, 2009; Chiticariu, 2010), and others (Fukumoto et al, 1998; Garigliano et al 1998; Humphrey et al, 1998). See (Bach and Badaskar, 2007) and (Konstantinova, 2014) for a comprehensive literature review of early relation extraction methods. All these approaches suffer from several drawbacks including: lack of portability of rules (Konstantinova, 2014), inherent limitation in scaling large number of rules (Reiss, 2008), ambiguity when resolving overlapping rules (Chiticariu, 2010), and the labor involved in hand-crafting and maintaining many rules.

Other approaches for relation extraction are based on supervised learning methods such as: logistic regression (Kambhatla, 2004), kernel-based (Zhao and Grishman, 2005; Bunescu and Mooney, 2006), Condition Random Field (Culotta, 2006). Semi-supervised methods such as (Brin et al, 1998; Agichtein and Gravano, 2000; and Etzioni, 2005). Most of these methods are focused on extracting only binary relations. (McDonald et al, 2005) presents a method for extracting higher-order relationships.

Since the formation of Information Extraction field, dating back to Message Understanding Conferences 1980's (MUC) sponsored by DARPA, relation extraction tasks have been exceptionally challenging. The state-of-the-art methods typically yield F1 performance between .60 and .70 on relation extraction datasets (Chiticariu, 2018). The methods reviewed above are considered traditional relation extraction methods, the critiqued paper (Lamurias et al) is based on current state-of-the-art methods based on deep learning using RNN/LSTM neural networks (Hochreiter, 1997) that have demonstrated F1 scores of .77 when extracting drug interaction relationships from biomedical text (W. Zheng et al, 2017). A common method used with these RNN/LSTM models is to include a word embedding layer trained using word2vec (Mikolov, 2013) or GloVe (Pennington, 2014) on a corpus of documents. Word embeddings have demonstrated significant improvement to performance because they can detect relationships between entities using unsupervised algorithms such as skip-gram and CBOW (continuous bag of words).

There are several well-known drawbacks with using word embedding including missing the underlying concepts and semantics behind words (Lucy and Gauthier, 2017), out-of-vocab problem (when words are not in the embedding space), inability to represent phrases (2 or more words), unable to distinguish multi-sense words (polysemy). See (Young et al, 2018) for a recent review of modern word embedding methods and limitations.

Motivated by the above success and limitations of word-embeddings, the critiqued paper proposes an additional embedding layer called: ontology-embedding. This approach attempts to leverage domain-specific ontologies as part of vector encoding process with the hypothesis that it will provide the down-stream layers with valuable information that is not directly observable through the training data. Specifically, the approach will focus on extracting ancestor concepts for each entity within the input text sequence and produce dense vectors representing based on the ancestor concepts.

Similar techniques of leveraging domain-specific information within a NN have demonstrated success such as (Xu, 2018; Li, 2016; Kong, 2013; and Deasigi, 2017). However, the critiqued paper asserts that no known attempt has been made in incorporating an ontology as part of an embedding layer to detect relations between entities. The paper proposes a model called BO-LSTM based on the RNN/LSTM architectures by (Y. Zhang, 2017) and (Y. Xu, 2015).

3. Methodology:

The key methodological aspects of the critiqued paper include:

1. Pre-processing pipeline and data preparation
2. Description of ontology embedding channels
3. Description of the proposed BO-LSTM model

3.1 Pre-processing pipeline and data preparation

Before the input text can be processed by the model, it must be pre-processed to encode the input sequence into vector representations. This pre-processing involves several well-known and common techniques in the field of NLP that include sentence tokenization, syntactic parsing, and detection of entities (i.e. entity chunking). The following steps describe this pre-processing pipeline:

1. Input text is tokenized into sentences using Spacy parser, a commonly used open source NLP python package).
2. Perform syntactic parsing and obtain Shortest-Dependency Path (SDP) structure for each package. This study again relies on Spacy NLP python package for this parsing.
3. Normalize entities using WordNet Hypernyms using method and tooling from (Ciaramita and Altun, 2006).
4. Resolve Concept Unique Identifier for each entity using an Ontology. For this study, two different ontologies were used. For the Drug-Drug Interaction benchmarks, the ChEBI Chemical Compound Ontology by (Hill et al, 2013) was used to train the ontology-embedding layer of the BO-LSTM model. For the Human-Phenotype benchmarks, the "Human Phenotype Ontology", developed by (Kuhler et al, 2017) was used for training the ontology-embedding layer.

A fuzzy matching approach is used to resolve unique concepts for each entity extracted from the input text sequence. This method for entity-to-concept fuzzy matching leverages the work by (Bhasuran et al, 2016) that demonstrated successful results. Although this match result is not guaranteed to be correct, it demonstrated robustness in similar biomedical entity recognition tasks. The fuzzy matching process performs an initial look-up using the preferred name of a concept within an ontology. The approach also performs a fuzzy matching of the synonyms for each concept and generates a Levenshtein distance (score) to decide which concept is the best match.

The training data is also prepared by performing class balancing between positive and negative examples bringing the ratio from 1:5.9 to 1:3.5 by eliminating entity pairs that are the same or when the only separator between entity pairs is punctuation (indicative of lists, enumerations, and abbreviations). Additionally, any entities that do not participate in a syntactic relationship (using SDP or syntactic parse structure) are excluded. These exclusion techniques for class balancing are based on work from (Abacha et al, 2015) and (Kim et al, 2015).

3.2 Ontology Embedding

A central aspect of the proposed BO-LSTM model relies on an “Ontology Embedding” layer that is trained together (vs pre-trained) with the rest of the BO-LSTM model for relationship extraction. The ontology embedding relies on the conversion of entities to a concept from an ontology (see Pre-Processing Pipeline above). For each pair of candidate entities in the training data, that may positively or negatively participate in a relationship, the ancestors for each entity is retrieved by traversing the ontology subsumption relations (i.e. IS-A). Each concept (including the ancestors and original concept) is encoded using the popular one-hot vector representation. This produces a set of sparse vectors that are sorted from generic to specific and then concatenated together to represent the input sequence to the ontology embedding layer.

The dimensionality of these sparse vectors is equivalent to the size of the ontology vocabulary (i.e. the number of unique concepts). This encoding representation does not scale when working with large ontologies (such as millions of concepts) due to the well-known curse of dimensionality problem. For this study, the vocabulary size was limited to 1,757 concepts from the ChEBI ontology. The total size of the ChEBI vocabulary is 109,000 concepts. This study was able to remove most all concepts by reducing the set to only the concepts that occurred within the datasets. In practice, this approach would yield a disproportionate number of out-of-vocabulary entities and therefore impact real-world performance of this model. This issue is further elaborated below (see Future work).

The critiqued paper also proposes an alternative to the concatenation of ancestor vectors. Instead of using all ancestors for both candidate concept (entity) pairs, only the common ancestors are encoded in the input sequence to the ontology embedding layer. The performance impact of these alternatives are compared and summarized below (see Contributions).

The purpose of the ontology embedding layer is to convert sparse concept vectors into dense vectors using an approach similar to word2vec (Mikolov, 2013) and GloVe (Pennington, 2014). The critiqued paper empirically tested different embedding dimensions of 50, 100, and 150 and concluded that 50 was the optimal dimension for this task. The sequence of encoded concept vectors is passed to an LSTM unit (Hochreiter, 1997) followed by a Max Pool layer. An illustration of the model architecture used for training the embedding space is shown in Figure 3 and Figure 1 of the critiqued paper. Training of the LSTM layer is accomplished through BPTT (Backpropagation Through Time) using Stochastic Gradient Descent (SGD) with Adam optimizer (Kingma, 2014) and a cross-entropy loss function. A dropout strategy of .5 was used as a regularization technique to avoid overfitting.

3.3 BO-LSTM Model

The critiqued paper uses a multi-channel NN architecture with each channel including multiple hidden layers that are aggregated using a dense layer ending with a Softmax classification layer. Figure 2 of the paper presents an illustration depicting this architecture. This multi-channel architecture approach is based on (Y. Xu, 2015) where they demonstrated improvement of model performance when aggregating multiple channels each representing different contextual information from the input sequence. The critiqued paper refers to their architecture as the "BO-LSTM" model. The model includes 4 channels. Two of these channels represent the proposed ontology-embedding layers (see Ontology Embedding above) one based on common ancestors between candidate concepts and the other a concatenation of ancestor concepts for each candidate pair. In addition to the ontology-embedding channels, two other channels are included in the architecture. One channel uses the Shortest-Dependency Path (SDP) to include words that interact with each other syntactically and the other channel uses WordNet Hypernym classes for normalization of entities. For additional details on how the SDP and WordNet classes are extracted see the Pre-processing pipeline above. The SDP embedding channel uses a pre-trained embedding produced by (Pyysalo, 2013). This pre-trained embedding is trained on 23 million PubMed's abstracts using word2vec (Mikolov, 2013). The WordNet embedding channel was trained together with the rest of the BO-LSTM model and uses a Bidirectional LSTM (Schuster, 1997; Graves, 2012; Baldi 1999) with Max Pooling configuration.

The output of each of these embedding channels are concatenated together and fed into a dense layer (fully connected NN) with a sigmoid activation function. The final layer of the BO-LSTM model uses the popular Softmax for classification. The classification layer is configured and trained to either provide a binary classification (i.e. whether there is a positive relationship between the candidate pair of entities) or a multinomial classification (i.e. describe the specific type of relationship between candidate pairs). Lamurias et al train these models using Keras with Tensorflow, a popular combination of python packages used to train deep learning models. Their work is made publicly available on Github.

4. Contributions:

Ontology Embedding Layer Improves Overall Performance when adapted to Deep Learning Models

To evaluate the significance of incorporating an Ontology Embedding layer within Neural Network model, Lamurias et al use the SemEval 2013 Drug-Drug-Interaction Challenge as a benchmark (Herrero-Zazo, 2013). This is a well-known challenge used by the NLP community to assess the performance of a model's ability to detect and classify relationships from biomedical text. Lamurias et al adapt their ontology embedding method to a similarly architected baseline model developed by (Y. Zhang et al, 2017). Prior to this experiment, the previous two highest performing models on the SemEval DDI challenge were (Y. Zhang et al, 2017) with F1 of .729 and (W. Zheng et al, 2017) with F1 of .773 both models were based on LSTM/RNN. The highest scoring model employed an Attention layer, the other leveraged classic word embedding layer.

After first replicating the recorded results using (Y. Zhang et al, 2017) model as a baseline, Lamurias et al adapted the ontology-embedding channels (both the common ancestors and concatenated ancestor channels) and demonstrated an improved F1 performance of .751 or a .022 improvement compared to the baseline F1 of .729. Results presented in Table 3 in (Lamurias et al, 2019) shows the original highest scoring model with F1 of .651 along with other high performing models and ending with this critiqued model (the second-best performing model) (Lamurias + Zhang 2017).

Additional observations between the baseline model and ontology-embedding enhanced model found that certain DDI relationships were only found by the enhanced model. This finding was further supported by performing a similar comparison using the BO-LSTM model and running it with and without the ontology-embedding channel. This observation supports the study's insight that ontologies contain important domain-specific information that is hidden and therefore not observable by deep learning models without the addition of an ontology embedding. Other noteworthy observations include the positive effect of the common ancestor's channel on precision and concatenated ancestors on recall measurements.

Thus, Lamurias et al's proposed ontology embedding method, which produces dense vectors using concept ancestor within a relevant ontology, is an effective method for enhancing performance of a Neural Network model trained to extract relationships.

5. Further research:

The proposed strategy of including an ontology embedding layer within a Neural Network architecture is promising and worthy of further investigation. However, there are several drawbacks and limitation with the proposed methods that require attention.

Chief among these limitations is the one-hot vector representation used for encoding concepts in an ontology has a major drawback. Large ontologies can scale to have millions of unique concepts and tens of millions of synonyms and relations. Using the sparse concept vector representation by Lamurias et al, would result in vectors with a million dimensions. Such a representation is not scalable and results in the well-documented curse of dimensionality problem. Lamurias et al avoided this problem by starting with a smaller ontology with only 109,000 concepts and further trimming the ontology vocabulary down to 1,757 concepts that occurred within the training and validation data. This approach would thus underperform in a real-world setting where the information to be processed is unknown and likely covering broad aspects of a given ontology. Additionally, the use of a restricted set of vocabulary within the ontology embedding layer will result in frequent occurrence of the Out-of-Vocab problem. This is a known limitation of embedding matrix where concepts do not have a vector position within the embedding and therefore are assigned a sentinel (or default) vector. This further degrades the performance of the model relative to observed validation accuracy. Recent research by (Rodriguez et al, 2018) proposes an eigenrepresentation alternate to one-hot vector encoding that may prove to be a viable alternative.

Lamurias et al suggest additional research into the selection of concept ancestors within the ontology. They suggest a better method for identifying only those ancestors with the highest information that can potentially be estimated by the probability of terms occurring within the context of the concept.

They also suggest the possibility of a semantic similarity measure to help find similarly related concepts. This limitation can be generalized further by describing the current method as only leveraging one kind of relation within the ontology, the subsumption (IS-A) relations. An ontology may contain many more valuable semantic information that is not being considered or made available to the down-stream layers of the model.

Other ideas for improvement include using linguistic heuristics such as discarding candidate pairs of entities that do not participate within a syntactic relationship, such as predicate-argument or subject-action-object structures. Another enhancement many involve using dynamic or on-the-fly embedding to handle out-of-vocab issues, such as the method proposed by (Herbelot, 2017) where an out-of-vocab word (or concept) is initiated with a vector sum of the context words and then refined as part of the training process with a high learning rate. Recent work with Zero-shot learning (ZSL) (L. Zhang, 2017) can help address the issue where the classification labels that a model must predict are continuously changing and therefore a model would become stale and require constant retraining. This is especially the case for industries that are evolving at a fast pace, such as biomedical and pharmaceutical industries, where new medical entities and relationships are being discovered. ZSL method would allow training a model that can handle the dynamic introduction of new classification labels without requiring retraining of the entire model. Attention mechanisms are another recent and relevant technique that help focus the model on specific areas of the input that are most predictive (see Chorowski, 2015; H. Xu, 2016). Lastly, other work to connect external memory such as Dynamic Memory Networks (DMN) (Kumar et al, 2016) and NTM (Neural Turing Machines) (Graves et al, 2016; Graves et al, 2014) may complement the ontology embedding layer by providing an additional source of domain-specific knowledge such as an ontology.

This suggests there are several areas of additional investigation that include the possibility of leveraging relevant research that has shown promise addressing similar problems.

REFERENCES

Lamurias, A., & Couto, F. M. (2019). Text mining for bioinformatics using biomedical literature. Encyclopedia of bioinformatics and computational biology, 1.

Chiticariu, L., Danilevsky, M., Li, Y., Reiss, F., & Zhu, H. (2018, June). SystemT: Declarative Text Understanding for Enterprise. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers) (pp. 76-83).

Rodríguez, P., Bautista, M. A., Gonzalez, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. Image and Vision Computing, 75, 21-31.

Xu, B., Shi, X., Zhao, Z., & Zheng, W. (2018). Leveraging biomedical resources in bi-lstm for drug-drug interaction extraction. IEEE Access, 6, 33432-33439.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.

Herbelot, A., & Baroni, M. (2017). High-risk learning: acquiring new word vectors from tiny data. arXiv preprint arXiv:1707.06556.

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. arXiv preprint arXiv:1705.11168.

Zheng, W., Lin, H., Luo, L., Zhao, Z., Li, Z., Zhang, Y., ... & Wang, J. (2017). An attention-based effective neural model for drug-drug interactions extraction. BMC bioinformatics, 18(1), 445.

Zhang, L., Xiang, T., & Gong, S. (2017). Learning a deep embedding model for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2021-2030).

Zhang, Y., Zheng, W., Lin, H., Wang, J., Yang, Z., & Dumontier, M. (2017). Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics, 34(5), 828-835.

Dasigi, P., Ammar, W., Dyer, C., & Hovy, E. (2017). Ontology-aware token embeddings for prepositional phrase attachment. arXiv preprint arXiv:1705.02925.

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., ... & Brudno, M. (2016). The human phenotype ontology in 2017. Nucleic acids research, 45(D1), D865-D876.

Bhasuran, B., Murugesan, G., Abdulkadhar, S., & Natarajan, J. (2016). Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. Journal of biomedical informatics, 64, 1-9.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... & Badia, A. P. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626), 471.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... & Socher, R. (2016, June). Ask me anything: Dynamic memory networks for natural language processing. In International conference on machine learning (pp. 1378-1387).

Li, Q., Li, T., & Chang, B. (2016). Learning word sense embeddings from word sense definitions. In Natural Language Understanding and Intelligent Applications (pp. 224-235). Springer, Cham.

Xu, H., & Saenko, K. (2016, October). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In European Conference on Computer Vision (pp. 451-466). Springer, Cham.

Abacha, A. B., Chowdhury, M. F. M., Karanasiou, A., Mrabet, Y., Lavelli, A., & Zweigenbaum, P. (2015). Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification. Journal of biomedical informatics, 58, 122-132.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. In Advances in neural information processing systems (pp. 577-585).

Kim, S., Liu, H., Yeganova, L., & Wilbur, W. J. (2015). Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. Journal of biomedical informatics, 55, 23-30.

Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., & Jin, Z. (2015). Classifying relations via long short term memory networks along shortest dependency paths. In proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1785-1794).

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. arXiv preprint arXiv:1410.5401.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Pennington, J., Socher, R., & 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).

Konstantinova, N. (2014, April). Review of relation extraction methods: What is new out there?. In International Conference on Analysis of Images, Social Networks and Texts (pp. 15-28). Springer, Cham.

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., & Declerck, T. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of biomedical informatics, 46(5), 914-920.

Hill, D. P., Adams, N., Bada, M., Batchelor, C., Berardini, T. Z., Dietze, H., ... & Hastings, J. (2013). Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC genomics, 14(1), 513.

Kong, X., Cao, B., & Yu, P. S. (2013, August). Multi-label classification by mining label and instance correlations from heterogeneous information networks. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 614-622). ACM.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., & Ananiadou, S. Distributional Semantics Resources for Biomedical Text Processing, 2013.

Graves, A. (2012). Supervised sequence labelling. In Supervised sequence labelling with recurrent neural networks (pp. 5-13). Springer, Berlin, Heidelberg.

Chiticariu, L., Li, Y., Raghavan, S., & Reiss, F. R. (2010, June). Enterprise information extraction: recent developments and open challenges. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 1257-1258). ACM.

Krishnamurthy, R., Li, Y., Raghavan, S., Reiss, F., Vaithyanathan, S., & Zhu, H. (2009). SystemT: a system for declarative information extraction. ACM SIGMOD Record, 37(4), 7-13.

Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Commun. ACM 51, 68–74 (2008)

Reiss, F., Raghavan, S., Krishnamurthy, R., Zhu, H., & Vaithyanathan, S. (2008, April). An algebraic approach to rule-based information extraction. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on (pp. 933-942). IEEE.

Bach, N., & Badaskar, S. (2007). A review of relation extraction. Literature review for Language and Statistics II, 2.

Bunescu, R., Mooney, R.: Subsequence kernels for relation extraction. In: Weiss, Y., Sch¨olkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems 18, pp. 171–178. MIT Press, Cambridge (2006)

Ciaramita, M., & Altun, Y. (2006, July). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 594-602). Association for Computational Linguistics.

Culotta, A., McCallum, A., Betz, J.: Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, pp. 296–303. Association for Computational Linguistics (2006)

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., ... & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, 165(1), 91-134.

Zhao, S., Grishman, R.: Extracting relations with integrated information using kernel methods. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 419 426. Association for Computational Linguistics, Morristown (2005)

Boguraev, B. K. (2004). Annotation-based finite state processing in a large-scale NLP architecture. Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003, 260, 61.

Kambhatla, N.: 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, Morristown (2004)

Agichtein, E., & Gravano, L. (2000, June). Snowball: Extracting relations from large plain-text collections. In Proceedings of the fifth ACM conference on Digital libraries (pp. 85-94). ACM.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. Bioinformatics, 15(11), 937-946.

Riloff, E., & Jones, R. (1999, July). Learning dictionaries for information extraction by multi-level bootstrapping. In AAAI/IAAI (pp. 474-479).

Brin, S.: Extracting patterns and relations from the world wide web. In: Proceedings of the First International Workshop on the Web and Databases, pp. 172–183 (1998)

Fukumoto, J., Masui, F., Shimohata, M., Sasaki, M.: Oki electric industry: description of the Oki system as used for MUC-7. In: Proceedings of the 7th Message Understanding Conference (1998)

Garigliano, R., Urbanowicz, A., Nettleton, D.J.: University of Durham: description of the LOLITA system as used in MUC-7. In: Proceedings of the 7th Message Understanding Conference (1998)

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., Wilks, Y.: University of Sheffield: description of the LaSIE-II system as used for MUC-7. In: Proceedings of MUC-7 (1998)

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673-2681.

Hearst, M. A. (1992, August). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics-Volume 2 (pp. 539-545). Association for Computational Linguistics.