A Constituency Parsing Tree based Method for Relation Extraction from Abstracts of Scholarly Publications

Ming Jiang, Jana Diesner
University of Illinois at Urbana-Champaign
{mjiang17,jdiesner}@illinois.edu

Abstract

We present a simple, rule-based method for extracting entity networks from the abstracts of scientific literature. By taking advantage of selected syntactic features of constituent parsing trees, our method automatically extracts and constructs graphs in which nodes represent text-based entities (in this case, noun phrases) and their relationships (in this case, verb phrases or preposition phrases). We use two benchmark datasets for evaluation and compare with previously presented results for these data . Our evaluation results show that the proposed method leads to accuracy rates that are comparable to or exceed the results achieved with state-of-the-art, learning-based methods in several cases.

1 Introduction

As a public and formal record of original contributions to knowledge, scientific literature is a critical resource that promotes the progress of science and technology in society. To help researchers to comprehend the large and growing amount of scientific literature, automated methods can be used to extract and organize information from corpora of publications, e.g., in terms of scientific key concepts and their relationships. Leveraging prior work that has achieved high accuracy for entity recognition (Lample et al., 2016; Habibi et al., 2017), in this paper, we focus on identifying extracting relationships between entities.

Overall, prior studies consider the task of relation extraction from two perspectives: 1) identifying if a relationship exists between a given pair of identified entities (Gábor et al., 2018), which is also the goal with this paper, and 2) further labeling or classifying the identified relationships (Luan et al., 2018a; Mintz et al., 2009).

Prior work has used different methods for relation extraction. Rule-based algorithms primarily rely on lexical patterns, such as word co-occurrences (Jenssen et al., 2001) and dependency templates (Fundel et al., 2006; Kilicoglu and Bergler, 2009; Romano et al., 2006). Supervised learning-based methods mainly use either feature engineering (Kambhatla, 2004; Chan and Roth, 2011) or kernel functions (Culotta and Sorensen, 2004; Zelenko et al., 2003; Bunescu and Mooney, 2005). Using supervised learning to detect the existence (and type) of relationships between concepts in scholarly publications may further require domain expertise for annotation. Accounting for the fact that humans tend to use their background knowledge to identify relations, Chan and Roth (2010) showed that using external knowledge, such as Wikipedia, improves the accuracy of relation extraction. Recently, deep learning models have also been used for supervised learning (Luan et al., 2018a, 2019). For example, Luan et al. (2019) developed a dynamic span graph framework based on sentence-level BiLSTM for multi-task information extraction. Since data annotation by humans is expensive, semi-supervised learning methods, such as the snowball system (Agichtein and Gravano, 2000), as well as unsupervised learning, such as clustering methods (Daelemans and Van den Bosch, 2005; Davidov and Rappoport, 2008), have also been explored in prior studies. Though remarkable contributions have been made, one remaining limitation with prior machine-learning based work is that these approaches may involve time and related costs for parameter optimization and labeling. Moreover, data-driven training methods may be limited by the domain specificity of learned models. Finally, algorithms trained on deep learning models lack interpretability.

To address the above-mentioned limitations, we

Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), pages 186–191
Hong Kong, November 4, 2019. ©2019 Association for Computational Linguistics

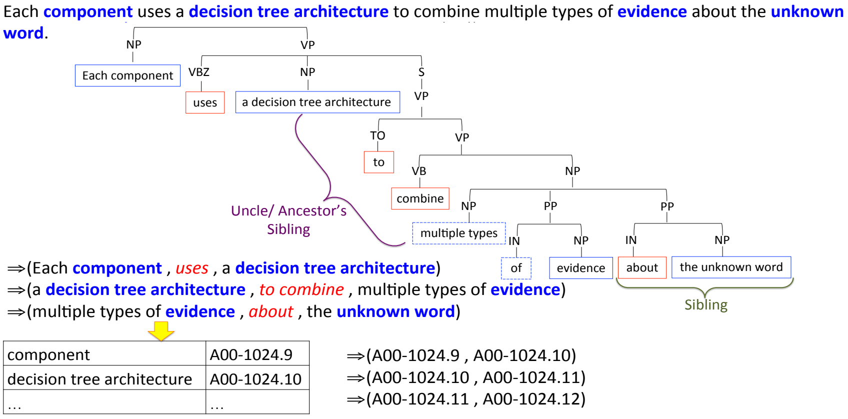
Each component uses a decision tree architecture to combine multiple types of evidence about the unknown word.
NP
VP
Each component
VBZ
NP
S
uses
a decision tree architecture
VP
TO
VP
to
VB
NP
combine
NP
PP
PP
Uncle/ Ancestor's Sibling
multiple types
IN
NP
IN
NP
of
evidence
about
the unknown word
Sibling
⇒(Each component , uses , a decision tree architecture)
⇒(a decision tree architecture , to combine , multiple types of evidence)
⇒(multiple types of evidence , about , the unknown word)
component | A00-1024.9
decision tree architecture | A00-1024.10
... | ...
⇒(A00-1024.9 , A00-1024.10)
⇒(A00-1024.10 , A00-1024.11)
⇒(A00-1024.11 , A00-1024.12)

Figure 1: An illustrative example of building a constituency-based concept network (CTN) for a sentence.

propose a rule-based concept network construction method. In the resulting networks, nodes represent annotated entities per document, and edges are constructed based on constituency parsing. Our approach is motivated by the observation that scientific literature typically use formal language with clear syntactic structure and comparatively fixed word order (e.g., term phrases). Inspired by prior work that defines a relationship as either an interaction or an association between two entities (Jurafsky, 2000), we take advantage of the structured information provided by constituency parse trees built for each sentence. More specifically, We capture two types of tuples: 1) (noun phrase, verb-based connection, noun phrase), and 2) (noun phrase, preposition-based connection, noun phrase). To avoid over-fitting on the given domain, our rules are generated on the basis of a context-free grammar.

We evaluate our method against two benchmark datasets that have been previously labeled for entities and relations. Our experimental results show that the proposed constituency-based concept networks achieve comparable accuracy to or can even outperform state-of-the-art, learning-based methods for identifying entity networks. We find that relationships of the type *used-for* and *part-of* are better captured by our approach than other types of relationships. Finally, we describe differences between domain-level concept networks.

2 Method

The construction of constituency-based concept network (CTN) has three stages: 1) data preprossessing, 2) the identification of nodes (we used entities given in annotated data), and of edges based on a constituency parsing tree, and 3) mapping entities from the constituency parsing tree to labeled entities. Figure 1 provides an illustrative example of the process of building a CTN for a sentence.

Data Pre-processing In this stage, we focus on two aspects. First, we segment each document into a set of sentences that are the input to constituency parsing. Since scientific texts may have some long sentences with complex sentence structures, we further segment sentences by using regular expressions. Second, we identify the annotated entities per sentence by extracting the labeled entity id and corresponding entity phrase, as well as the entities' index in the sentence.

Node and Edge Identification To generate a parsing tree for each sentence, we use the AllenNLP constituency parsing (Joshi et al., 2018) As shown in Figure 1, we extract noun phrases at the lowest layer (i.e., children are noun-based end-nodes) as candidates of CTN nodes. We made this decision to capture unique and specific (to the sub-field of science) noun phrases from scientific literature. For linking nodes, we capture keywords that occur between two adjacent node candidates in the parse tree. According to the parse tree structure, these keywords usually occur in two types of positions: 1) the sibling of a CTN node candidate, and 2) the uncle or the ancestor's sibling of a potential CTN node. Based on node candidates and connecting keywords, we generate an edge if two node candidates are linked by a connecting keyword. One specific issue in this process is the of-phrase: according to our empirical observation, we believe that of-phrases (e.g., "computational model of discourse") often represent a single concept. Based on this assumption, we merge

	SciERC	SemEval18
#entities	8089	7482
#relations	4716	1595
#relations/doc	9.4	3.2
cross-sentence relations	yes	no

Table 1: Data statistics.

edge candidates connected by the keyword "of" as a single CTN node, and remove the original edges.

Entity Mapping We remove nodes that had been identified by the constituency parsing tree process described above, but are not labeled as entities in the ground truth data. As shown in Figure 1, the final output of CTN is a set of nodes with ids, and edges.

3 Experiments

3.1 Experimental Setup

Data We perform experiments on two publicly available datasets where humans annotated scientific entities and relations from abstracts of scientific publication. Table 1 provides a brief summary of both datasets. ***SemEval18 Dataset*** has 500 abstracts prepared by Gábor et al. (2018) for the shared task 7 (subtask 2) of SemEval 2018. All abstracts in this dataset are from published papers in the field of computational linguistics. The annotated relations are divided into six types of semantic relationships between scientific concepts. The ***SciERC Dataset*** was provided by Luan et al. (2018a). This dataset has 500 scientific abstracts from 12 AI conference/workshop proceedings that cover five research areas: 1) artificial intelligence (AI), 2) natural language processing (NLP), 3) speech, 4) machine learning (ML), and 5) computer vision (CV).

Baselines For the SemEval data, we compared the results from our network construction method (CTN) with the official baseline, which was generated by using a memory-based k-nearest neighbor (k-nn) search (Gábor et al., 2018). The top three reported submissions in the SemEval leaderboard were: UWNLP (Luan et al., 2018b), ETH-DS3Labl (Rotsztejn et al., 2018), and SIRIUS-LTG-UiO (Nooralahzadeh et al., 2018). For SciERC, we compared our method with two state-of-the-art (SOTA) systems: SciIE (Luan et al., 2018a) and DyGIE (Luan et al., 2019). The original outputs from SciIE and DyGIE also identified and categorized the type and direction of relations, and the boundary of entities. Since we do not identify these elements, we also do not consider them for comparison.

	Dev			Test		
	P	**R**	**F1**	**P**	**R**	**F1**
SciIE	62.0	47.7	53.9	66.4	46.7	54.9
DyGIE	55.0	**48.6**	51.6	63.3	**52.2**	57.2
CTN (ours)	**73.4**	47.3	**57.5**	**75.4**	46.5	**57.5**

Table 2: Comparison with previous methods for relation extraction on SciERC dataset.

3.2 Results

Extraction Performance Table 2 and Table 3 compare our concept network CTN with baselines for SciERC and SemEval18, respectively. Compared to SOTA on SciERC, our approach outperformed the best previously reported results on both the development and testing data. Specifically, CTN achieved a noticeable improvement in precision (we achieved ∼74%) compared to prior methods, which benefits our F1 value. Further comparing each method's performance on the development data versus testing data, we observe that our rule-based CTN produces more stable or consistent results than the considered, prior, learning-based methods.

For SemEval18, we find that CTN outperformed the baseline by ∼15% (our F1 score was 41.6%), and the third best reported prior result, but could not come close to the top two prior results. From the presented results, we conclude that our rule-based concept network approach can serve as a strong baseline method for identifying related entity pairs in scientific texts.

Ablation Study In order to explore the isolated contribution of each rule considered for adding edges in the CTN, we conduct an ablation analysis

	P	**R**	**F1**
UWNLP	-	-	**50.0**
ETH-DS3Lab	-	-	48.8
SIRIUS-LTG-UiO	-	-	37.4
SemEval Baseline	-	-	26.8
CTN (ours)	33.6	54.8	41.6

Table 3: Comparison with previous methods for relation extraction on SemEval18 dataset.

	Dev			Test		
	P	**R**	**F1**	**P**	**R**	**F1**
CTN (ours)	**73.4**	**47.3**	**57.5**	**75.4**	**46.5**	**57.5**
– long sentence segmentation	73.4	47.3	57.5	75.4	46.5	57.5
– of-phrase (merge+remove)	75.1	44.4	55.8	76.0	43.8	55.6
– of-phrase (merge)	74.6	40.7	52.6	76.6	41.3	53.6
– entity id positioning	64.0	31.2	42.0	72.9	32.6	45.1
– all above	64.1	29.0	39.9	73.1	31.0	43.5

Table 4: Ablation study of isolated contribution of each rule.

SciERC			**SemEval18**		
	TP%	Total		TP%	Total
USED-FOR	57.22%	533	PART_WHOLE	69.51%	82
PART-OF	50.79%	63	USAGE	59.77%	174
FEATURE-OF	49.15%	59	RESULT	50.00%	16
EVALUATE-FOR	43.96%	91	COMPARE	36.83%	19
HYPONYM-OF	40.30%	67	MODEL-FEATURE	34.25%	73
COMPARE	36.84%	38	TOPIC	0.00%	3
CONJUNCTION	4.88%	123	-	-	-

Table 5: Accuracy (TP = true positives) per relationship type on testing data.

by removing each rule from the network construction process and measuring the overlap between predicted connections and the ground truth. Table 4 shows the results. We observe that the rule for mapping entity ids for a potentially connected entity pair has the highest isolated impact. By further looking into the ground truth data, we find that an entity phrase can be labeled by multiple different ids, e.g., when the related phrase repeatedly appears in a document. Therefore, without a mapping step, CTN would be prone to considering and linking such entities as different nodes. Coming back to of-phrases, we find that not considering edges between the preposition "of" (i.e., (node 1, "of", node 2)) leads to a decrease in recall, which further indicates our initial recommendation that of phrases should be merged to represent a single (scientific) concept.

Relation Type Sensitivity Table 5 shows the ability of CTN to identify each type of relationship that is labeled in the ground truth data. CTN does not actually predict relationship type, we only retroactively compute the accuracy rate per link type. We observe that *usage (used-for)* and *part_whole (part-of)* are identified with the highest accuracy. Note that these two categories are also the most frequently occurring ones in the ground truth data. On the other hand, we find that relationship type of conjunction and topic association result in the lowest accuracy. This might be because we do not consider conjunction words as keywords that indicate relationships. To understand the comparatively low performance for topic relationships, we further looked into the context of the actual entity pairs. Doing so, we found that all three instance of this link type were expressed by an of-phrase, where we consider the whole phrase as a single concept. For example, in the expression "qualitative analysis of results", the entity "qualitative analysis" is annotated with a topic relationship with the entity "results".

Network Analysis To understand the characteristics of the network data that were constructed with the proposed method in more depth, we further built a corpus-level CTN for all texts per domain in the SciERC testing data. Figure 3 shows two illustrative examples of the networks for the CV and ML domain, respectively. The node name represents the extracted entity phrases, node size represents the weighted degree centrality, and node colors denote membership in components.

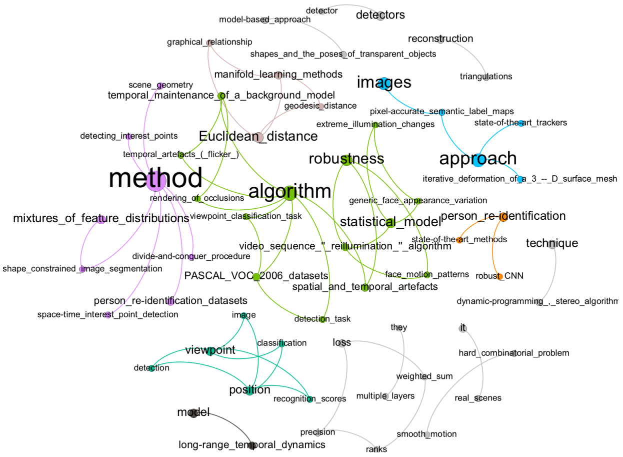
detector
detectors
model-based_approach
graphical_relationship
reconstruction
shapes_and_the_poses_of_transparent_objects
manifold_learning_methods
triangulations
images
scene_geometry
temporal_maintenance_of_a_background_model
geodesic_distance
pixel-accurate_semantic_label_maps
extreme_illumination_changes
state-of-the-art_trackers
detecting_interest_points
Euclidean_distance
temporal_artefacts_(_flicker_)
robustness
approach
method
iterative_deformation_of_a_3_--_D_surface_mesh
algorithm
rendering_of_occlusions
generic_face_appearance_variation
mixtures_of_feature_distributions
viewpoint_classification_task
statistical_model
person_re-identification
state-of-the-art_methods
video_sequence_"_reillumination_"_algorithm
technique
divide-and-conquer_procedure
shape_constrained_image_segmentation
PASCAL_VOC_2006_datasets
face_motion_patterns
robust_CNN
spatial_and_temporal_artefacts
person_re-identification_datasets
dynamic-programming_,_stereo_algorithm
space-time_interest_point_detection
image
detection_task
they
it
classification
loss
viewpoint
hard_combinatorial_problem
detection
weighted_sum
position
recognition_scores
multiple_layers
real_scenes
model
precision
smooth_motion
long-range_temporal_dynamics
ranks

Figure 2: CTN of abstracts from the CV domain.

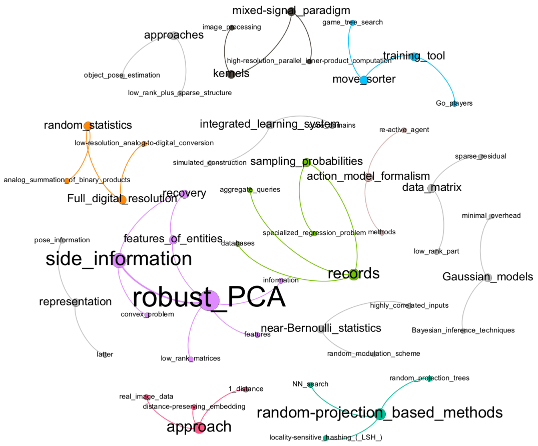

Figure 3: CTN of abstracts from the ML domain.

We find that the most central (in terms of degree) nodes from the CV corpus mainly represent general scientific concepts, such as "method", "algorithm" and "approach", while in the ML corpus, key nodes represent domain-specific terms such as "robust PCA" and "side information". Comparatively, the size of the CTN from the CV domain is larger than that from the ML domain, which might be due to different numbers of abstracts in each domain.

4 Conclusions

In this paper, we have proposed and evaluated a rule-based network construction method that leverages constituency parsing to extract relations between entities in scientific texts. Our method does not require machine learning or domain knowledge. Experiments on two benchmark datasets show that the proposed CTN achieve comparable performance with state-of-the-art learning-based methods in multiple cases. Even though our method could not outperform the two best performing systems built for one of the considered datasets, our results suggest that the demonstrated approach can work as a baseline method for relation extraction. In addition, we find that entities with a relationship of *used-for* and *part-of* are more likely to be connected in our network. Based on the corpus-level CTN, we further saw that key nodes in the networks based on the CV corpus are mainly general scientific terms, while for the abstracts from the ML domain, key nodes represent domain-specific terms. To improve the construction of CTN in the future, we plan to consider cross-sentence link formation and link label detection. Finally, the rules used to build CTN can further support the development of learning-based algorithms for relation extraction.

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number, 2015-ST-061-CIRC01. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM Conference on Digital Libraries*, pages 85–94. ACM.

Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731. Association for Computational Linguistics.

Yee Seng Chan and Dan Roth. 2010. Exploiting background knowledge for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 152–160. Association for Computational Linguistics.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 551–560. Association for Computational Linguistics.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 423–429, Barcelona, Spain.

Walter Daelemans and Antal Van den Bosch. 2005. *Memory-based Language Processing*. Cambridge University Press.

Dmitry Davidov and Ari Rappoport. 2008. Classification of semantic relationships between nominals using pattern clusters. In *Proceedings of ACL-08: HLT*, pages 227–235, Columbus, Ohio. Association for Computational Linguistics.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2006. Relexrelation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski, and Eivind Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1199, Melbourne, Australia. Association for Computational Linguistics.

Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 178–181, Barcelona, Spain. Association for Computational Linguistics.

Halil Kilicoglu and Sabine Bergler. 2009. Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 119–127.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018a. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2018b. The uwnlp system at semeval-2018 task 7: Neural relation extraction model with selectively incorporated concept embeddings. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 788–792.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. 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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Sirius-ltg-uio at semeval-2018 task 7: Convolutional neural networks with shortest dependency paths for semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 805–810.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang. 2018. Eth-ds3lab at semeval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 689–696.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3(Feb):1083–1106.