E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 608–614

PARSING STRUKTUR PARAGRAF BERBASIS NEURAL NETWORK

Agung Prasetya[1)], Taufiq Agung Cahyono[2)]
[1, 2)]Informatika, Fakultas Sains dan Teknologi, Universitas Bhinneka PGRI
Jl. Mayor Sujadi Timur, Kec. Kedungwaru, Kab. Tulungagung
e-mail: agung@ubhi.ac.id[1)], taufiq@ubhi.ac.id[2)]

ABSTRAK

Parsing paragraf memiliki peran penting dalam perkembangan kecerdasan buatan. Parsing menjadi langkah awal untuk menalar paragraf agar bisa dimengerti oleh mesin. Keefektifan metode parsing paragraf bergantung pada bagaimana mendekomposisikan teks ke segmen teks. Proses segmentasi tanpa memperhitungkan struktur semantik dari suatu paragraf akan menghasilkan struktur yang tidak sinkron dengan makna sebenarnya. Untuk mengatasi masalah ini, penelitian ini mengusulkan penerapan metode berbasis recursive neural network (RvNN). Metode ini berupaya mendapatkan binary tree terbaik yang merepresentasikan struktur paragraf. Metode usulan diterapkan untuk menyelesaikan paragraf-paragraf sederhana yaitu soal cerita anak. Hasil uji coba menunjukkan bahwa metode usulan dapat memparsing paragraf dengan tingkat akurasi sebesar 0.9. Metode usulan juga lebih efisien karena tidak perlu membuat repositori kerangka struktur.

***Kata Kunci**: paragraf, parsing, recursive neural network, semantik, binary tree*

ABSTRACT

Paragraph parsing has an important role in the development of artificial intelligence. Parsing is the first step to reasoning paragraphs so that they can be understood by machines. The effectiveness of the paragraph position parsing method on how to decompose text into text segments. The segmentation process without taking into account the semantic structure of a paragraph will result in a structure that is not in sync with the actual meaning. To overcome this problem, this study proposes recursive neural network (RvNN) based method. This method tries to get the best binary tree that represents paragraph structure. The proposed method is applied to complete simple paragraphs about children's stories. The test results show that the proposed method can parse paragraphs with an accuracy rate of 0.9. The proposed method is also more efficient because there is no need to build the repository of paragraph structure..

Keywords: *paragraph, parsing, recursive neural network, semantic, binary tree*

I. PENDAHULUAN

PARAGRAF memiliki peranan dalam perkembangan bidang kecerdasan buatan. Paragraf dapat dijadikan sebagai acuan untuk pembuatan mesin yang dapat mengerti bahasa alami manusia [1]. Mesin tersebut harus dapat menalar input berupa teks narasi.

Masalah utama pada paragraf adalah bagaimana mendeteksi struktur paragraf yang nantinya akan digunakan pada penalaran [1], [2]. Metode-metode berbasis kerangka mendeteksi jenis struktur paragraf dengan cara pencocokan teks terhadap template-template [3]–[5]. Metode-metode berbasis kerangka harus menyediakan kumpulan template kerangka struktur paragraf. Metode-metode berbasis kerangka kurang efektif ketika menerima paragraf yang tidak ada pada repositori template. Kesulitan ini diperbaiki oleh metode-metode berbasis parsing [3], [6]–[8]. Metode-metode berbasis parsing mendekomposisikan teks ke sejumlah segmen-segmen. Segmen-segmen tersebut selanjutnya diklasifikasikan menggunakan model klasifikasi teks. Metode-metode berbasis parsing lebih efisien dibandingkan metode berbasis kerangka karena tidak memerlukan pembentukan repositori template.

Kefektifan metode-metode berbasis parsing tergantung pada keefektifan proses dekomposisi teks ke segmen. Dekomposisi tersebut tidak bisa ditentukan hanya dengan pembatasan teks menggunakan dua pembatas kalimat saja. Pembatasan teks harus memperhitungkan struktur semantik. Tanpa memperhitungkan struktur semantik, dekomposisi teks akan menghasilkan segmen yang maknanya tidak menunjuk ke operasi aritmetik. Pada kasus paragraf, proses deteksi struktur paragraf akan menghasilkan struktur paragraf yang tidak sinkron dengan makna paragraf.

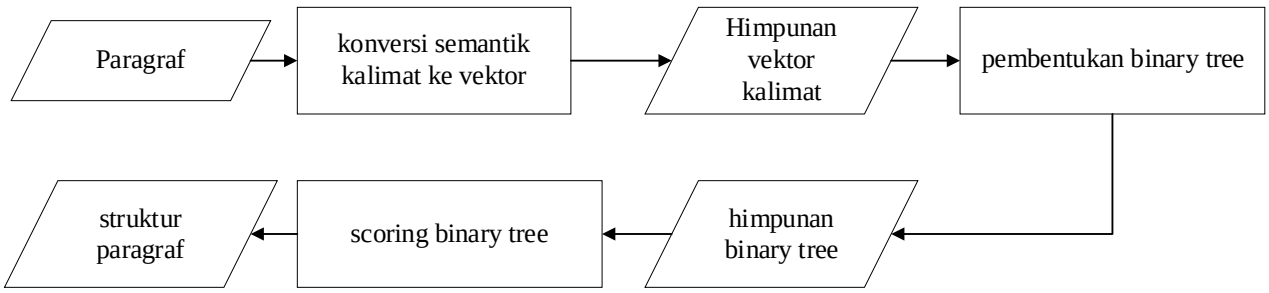
Berdasarkan permasalahan diatas, penelitian ini mengusulkan metode untuk memparsing struktur semantik paragraf. Proses parsing dimodelkan sebagai masalah prediksi struktur. Penelitian ini mengusulkan penerapan *Recursive Neural Network* (RvNN) [9], [10] untuk menyelesaikan masalah prediksi struktur semantik paragraf. Model RvNN sebelumnya telah digunakan pada parsing struktur suatu gambar.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 608–614

II. METODE

Metode usulan pada penelitian ini memodelkan struktur semantik paragraf menjadi *binary tree*. Masing-masing node merepresentasikan segmen-segmen teks. Metode usulan tersusun atas tiga fase, yaitu konversi semantik kalimat ke vektor, pembentukan *binary tree* dan scoring *binary tree*. Luaran fase pertama menjadi masukan fase kedua. Luaran fase kedua menjadi masukan fase ketiga. Gambar 1 menunjukkan diagram dari metode usulan



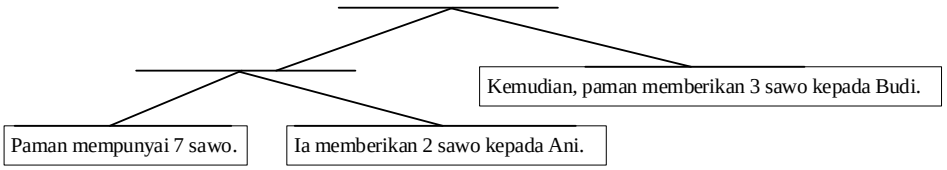
Gambar. 1. Diagram metode parsing paragraf usulan

Fase pertama membagi terlebih dahulu paragraf ke sejumlah kalimat. Pembagian ini dilakukan dengan menerapkan *regular expression* (regex) yang diambil dari [11]. Regex untuk membagi paragraf ke kalimat ditunjukkan pada gambar 2. Selanjutnya, kalimat-kalimat hasil dari penerapan regex dikonversi ke vektor menggunakan metode sentence2vector [12], [13]. Metode ini menggunakan *recursive* autoencoder.

```
[^.!?\s][^.!?]*(?:[.!?](?!['"]?\s|$)[^.!?]*)*[.!?]?['"]?(?=\s|$)
```

Gambar. 2. Regex untuk mengidentifikasi kalimat pada sebuah paragraf

Fase kedua adalah pembentukan *binary tree*. Untuk satu paragraf, fase ini membentuk sekumpulan *binary tree*. Seperti yang telah diuraikan, node-node pada *binary tree* adalah kalimat-kalimat dari fase pertama. Diberikan n-kalimat, jumlah *binary tree* yang mungkin bisa dibentuk dari satu paragraf adalah sejumlah $\frac{(2n)!}{(n+1)!n!}$. Contoh sebuah *binary tree* dari sebuah paragraf sederhana "Paman mempunyai 7 sawo. Ia memberikan 2 sawo kepada Ani. Kemudian, paman memberikan 3 sawo kepada Budi" ditunjukkan pada gambar 3.



Gambar. 2. Contoh binary tree dari sebuah paragraf sederhana

Fase ketiga adalah memberikan score pada setiap binary tree dari hasil proses fase kedua. Score *binary tree* dihitung dengan fungsi prediksi f. Diberikan himpunan kalimat soal cerita x dan parameter θ , fungsi prediksi f yaitu:

$$f(x,\theta) = \arg\max_{t \in T(x)} RvNN(t,\theta)$$

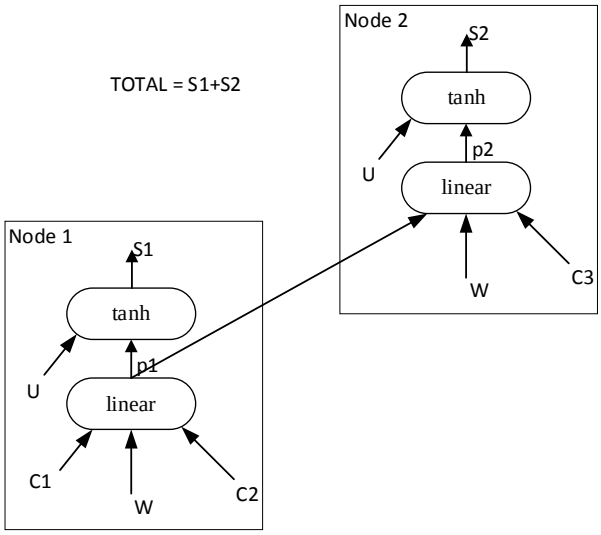
yang mana T adalah himpunan *binary tree* dari fase kedua.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 608–614

Fungsi prediksi f menggunakan model *Recursive* Neural Network ($RvNN$). RvNN menggunakan dua buah fungsi aktivasi per node pada *binary tree*. Kedua fungsi aktivasi tersebut yaitu fungsi linear dan non-linear tanh. Fungsi linear mengambil masukan dua buah vektor baik dari luaran child node atau vektor kalimat dari hasil fase pertama. Gambar 4 mengilustrasikan struktur RvNN.

Skor total pada RvNN dihitung secara bottom-up. Proses perhitungan diawali dari node yang terdalam hingga node teratas (root node). Masukan sebuah node akan mengambil dua buah masukan, baik berupa luaran node dibawahnya ataupun vektor kalimat. Metode usulan menggunakan parameter-parameter fungsi aktivasi yang sama di setiap node. Parameter θ pada fungsi prediksi f adalah bobot W pada fungsi linear dan bobot U pada fungsi non-linear $tanh$. Penentuan kedua parameter ini dilakukan dengan cara memilih nilai-nilai yang dapat menghasilkan



Gambar. 4. Struktur RvNN

skor total yang besar untuk *binary tree* yang benar dan skor total yang kecil untuk *binary tree* yang salah.

Permasalahan dalam penentuan parameter θ diselesaikan melalui supervised learning. Terkait hal ini, sebuah dataset dibentuk yang memuat sejumlah pasangan soal cerita dan *binary tree*. Mengadopsi fungsi obyektif pada[14], [15], fungsi obyektif penentuan parameter θ yaitu:

$$J(\theta) = \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{l(t_i, t) + RvNN(t, \theta)\} - RvNN(t_i, \theta)$$

yang mana λ adalah regularizer. Fungsi l mengukur perbedaan dua buah struktur *binary tree*. Metode usulan ini menggunakan fungsi structured loss margin. Fungsi ini akan menghitung perbedaan *binary tree* yang benar yang ada pada dataset dengan suatu *binary tree* t. Untuk setiap iterasi, fungsi obyektif diupayakan untuk diminimalkan.

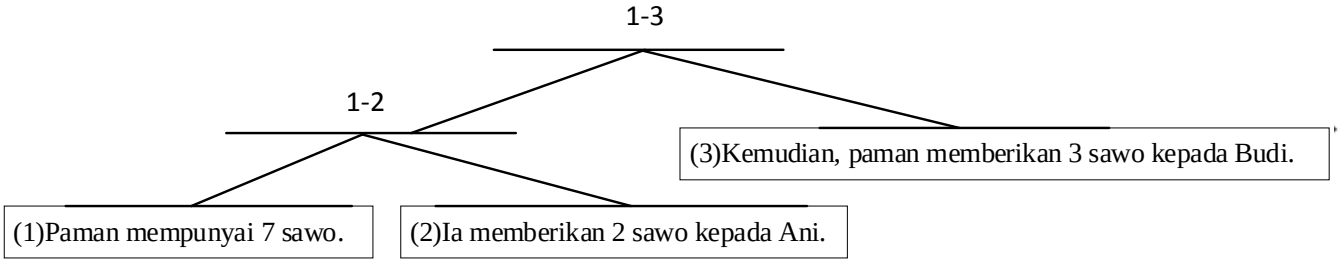
Fungsi structured loss margin l pada metode usulan mengadopsi [16], [17]. Fungsi structured loss margin l bernilai 0 apabila dua buah *binary tree* memiliki struktur yang sama $l(t_i, t) > 0, \forall t, t \in T(x_i) \backslash t_i$ dan $l(t_i, t_i) = 0$. Menggunakan cara yang sama pada [16], [17], metode usulan menghitung berapa banyak span yang berbeda diantara dua *binary tree*. Span didefinisikan sebagai pasangan index yang merepresentasikan posisi kalimat terkiri dan terkanan dibawah sebuah non-terminal node. Ilustrasi span dari suatu struktur paragraf ditunjukkan pada gambar 5.

Perbedaan span akan bernilai 1. Span yang sama bernilai 0. Fungsi l menghitung keseluruhan span yang berbeda. Persamaan untuk fungsi structured loss margin l yaitu:

$$l(t_i, t) = \sum_{d \in R(t)} 1\{d \notin R(t_i)\}$$

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 608–614



Gambar. 5. Index span pada sebuah binary tree dari suatu paragraf

yang mana $R(t_i)$ adalah himpunan *span* di *non-terminal node* pada suatu *binary tree* t_i dan d adalah *non-terminal node.*

III. HASIL UJI COBA DAN PEMBAHASAN

A. Dataset

Paragraf yang digunakan sebagai ujicoba pada penelitian ini adalah soal cerita matematika sekolah dasar. Soal cerita yang menjadi dataset diambil dari http://bse.kemdikbud.go.id/ . Soal cerita yang digunakan sebagai uji coba hanya soal-soal yang memuat dua buah aritmetika dasar (pengurangan, penambahan, perkailan dan pembagian). Jumlah soal cerita untuk pembentukan dataset ditunjukkan pada tabel 1.

TABEL I
STATISTIK DATASET SOAL CERITA
.

Jenis Karakteristik Dataset	Nilai
Jumlah total soal	1200
Jumlah soal campuran	308
Jumlah soal non-campuran	904
Jumlah soal cerita dengan jumlah kalimat <3	167

Masing-masing soal yang dikumpulkan selanjutnya disusun binary tree menggunakan aplikasi Rhetorical Structure versi 3.0 [18]. Soal-soal yang menjadi dataset untuk parser struktur paragraf dipilih soal yang tersusun atas lebih dari 3 kalimat. Hal ini dilakukan untuk menghindari adanya paragraf yang tidak bermakna dan mneghindari paragraf yang strukturnya tidak membentuk binary tree. Ada beberapa pola binary tree yang dapat terbentuk. Pola-pola tersebut ditunjukkan pada tabel 2.

TABEL 2
JENIS POLA DAN JUMLAH PARAGRAF
.

Jenis Pola	Jumlah Pola
Jumlah pola 1	1200
Jumlah pola 2	308
Jumlah pola 3	904
Jumlah pola 4	167
Jumlah pola 5	230

Contoh paragraf yang tergolong pola 1 dan 2 ditunjukkan pada gambar 6 dan gambar 7

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 608–614

B. Hasil Uji Coba

Metode usulan harus mengubah kalimat ke vektor. Uji coba dilakukan untuk mengetahui dimensi vektor yang tepat untuk metode parsing. Dimensi vektor yang diujicobakan adalah 50, 100 dan 200. Untuk mengurangi bias, pengujian dilakukan dengan k-fold cross validation dengan $k = 6$. Baik data latih maupun data uji diperoleh dengan membagi masing-masing pola sebanyak k. Penentuan data ke setiap fold dilakukan secara acak. Untuk setiap 1 fold, kelima fold digabungkan sebagai data latih dan 1 fold sebagai data uji.

Data latih digunakan untuk menentukan bobot regularizer λ dan nilai penalti structure loss κ. Hyperparameter ini ditentukan menggunakan batch subgradient descent. Untuk setiap iterasi ke-i, nilai rate learning ditentukan berdasarkan rule constatnt step length $\alpha_i = \frac{\gamma}{\|g_i\|_2}$ yang mana g adalah parameter gradient dan γ merupakan koefisien. Nilai koefisien ditentukan sebesar 0.8. Tabel 3 menunjukkan hyperparameter yang diujicobakan.

TABEL 3
HYPERPARAMETER YANG DIUJICOBAKAN

Parameter	Nilai
Bobot γ	1e-6, 1e-5,1e-4,1e-3,5e-3
Penalti structure loss κ	1.0,0.9,0.8,0.7,0.6
Dimensi vektor kalimat	50,100 dan 200

```
    </node>
   <node mathoperation="nooperation" type="leaf" id="3">berapa jumlah telur yang ani beli</node>
  </node>
 </node>
```

Gambar. 7. Contoh paragraf yang termasuk pola 3

Uji coba awal digunakan untuk mengukur pengaruh regularizer λ terhadap tingkat keakuratan metode usulan. Hasil akurasi untuk sejumlah nilai λ ditunjukkan pada tabel 4. Untuk uji coba ini, dimensi vektor kalimat ditetapkan sebesar 0.8 dan nilai penalti structure loss sebesar 50.

TABEL 4
AKURASI METODE USULAN TERHADAP NILAI REGULARIZER

Regularizer λ	Akurasi Fold0	Akurasi Fold 1	Akurasi Fold 2	Akurasi Fold 3	Akurasi Fold 4	Rata-rata Akurasi
1e-6	0.85	0.87	0.85	0.83	0.76	0.84
1e-5	0.88	0.87	0.84	0.85	0.77	0.84
1e-4	0.88	0.89	0.89	0.89	0.80	0.86
1e-3	0.84	0.88	0.85	0.84	0.79	0.85
5e-3	0.87	0.86	0.86	0.86	0.76	0.85

Berdasarkan tabel 4 akurasi terbaik untuk metode usulan diperoleh untuk nilai regularizer sebesar 1e-4. Meskipun demikian, perbedaan akurasi tidak menunjukkan nilai yang signifikan.

Uji coba selanjutnya digunakan untuk mengetahui nilai structure loss κ yang dapat menghasilkan nilai akurasi terbaik. Untuk uji coba ini, regularizer ditentukan dari nilai terbaik pada uji coba pertama yaitu 1e-4. Hasil uji coba pengaruh structure loss κ ditunjukkan pada tabel 5.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 608–614

TABEL 5
AKURASI METODE USULAN TERHADAP STRUCTURE LOSS

Penalti structure loss λ	Akurasi Fold0	Akurasi Fold 1	Akurasi Fold 2	Akurasi Fold 3	Akurasi Fold 4	Rata-rata Akurasi
0.9	0.86	0.78	0.85	0.83	0.76	0.84
0.8	0.89	0.76	0.84	0.85	0.77	0.84
0.7	0.9	0.89	0.9	0.89	0.90	0.89
0.6	0.87	0.88	0.86	0.86	0.76	0.85

Berdasarkan tabel 5 akurasi terbaik dari metode usulan diperoleh untuk nilai penalti structure loss sebesar 0.7. Sama halnya dengan uji coba pertama, perbedaan akurasi tidak menunjukkan nilai yang signifikan.

Uji coba ketiga bertujuan untuk mengetahui dimensi vektor kalimat yang dapat menghasilkan nilai akurasi terbaik. Untuk uji coba ini, regularizer ditentukan dari nilai terbaik pada uji coba pertama yaitu 1e-4 dan penalti structure loss sebesar 0.7. Hasil uji coba pengaruh dimensi vektor kalimat ditunjukkan pada tabel 6.

TABEL 6
AKURASI METODE USULAN DIMENSI VEKTOR KALIMAT

Penalti structure loss λ	Akurasi Fold0	Akurasi Fold 1	Akurasi Fold 2	Akurasi Fold 3	Akurasi Fold 4	Rata-rata Akurasi
50	0.86	0.78	0.85	0.83	0.76	0.82
100	0.89	0.76	0.84	0.85	0.77	0.83
200	0.93	0.91	0.92	0.91	0.9	0.92

Berdasarkan tabel 6 akurasi terbaik dari metode usulan diperoleh untuk vektor kalimat sebesar 200. Berbeda dengan uji coba-uji coba sebelumnya, perbedaan akurasi menunjukkan nilai yang signifikan. Hal ini menunjukkan adanya pengaruh yang kuat dalam pemilihan dimensi vektor kalimat dengan tingkat akurasi.

C. Pembahasan

Berdasarkan hasil uji coba, dimensi vektor kalimat sangat berpengaruh terhadap akurasi metode usulan berbasis RvNN. Tingkat akurasi semakin baik ketika dimensi vektor semakin besar. Peningkatan ini disebabkan karena dimensi yang besar dapat merepresentasikan secara detil paragraf. Berbeda dengan vektor berdimensi besar, vektor berdimensi kecil kurang memiliki tingkat detil pada medan vektor.

IV. KESIMPULAN

Uji coba menunjukkan bahwa metode usulan berbasis RvNN mampu memparsing struktur paragraf dengan akurasi sebesar 0.864. Nilai akurasi tesebut menggunakan penalti structure loss 0,8 dan regularizer 1e-4. Hasil akurasi terbaik diperoleh untuk dimensi kalimat sebesar 200. Hasil uji coba menunjukkan pula bahwa faktor yang berpengaruh adalah dimensi vektor kalimat. Penelitian selanjutnya perlu mempertimbangkan untuk metode pembentukan binary tree yang lebih efisien. Penelitian selanjutnya juga dapat mempertimbangkan penggabungan fungsi obyektif per node dengan fungsi obyektif keseluruhan *binary tree* melalui *joint learning*. Penelitian selanjutnya perlu menyelesaikan paragraf-paragraf dengan struktur yang lebih kompleks.

DAFTAR PUSTAKA

[1] P. Clark and O. Etzioni, “My Computer Is an Honor Student — But How Intelligent Is It? Standardized Tests as a Measure of AI,” *ojs.aaai.org*, 2016, Accessed: Jun. 18, 2022. [Online]. Available: www.allenai.org

[2] P. C.- AAAI and undefined 2015, “Elementary school science and math tests as a driver for AI: Take the aristo challenge!,” *Citeseer*, Accessed: Jun. 18, 2022. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.697.9514&rep=rep1&type=pdf

[3] L. Zhou, S. Dai, and L. Chen, “Learn to Solve Algebra Word Problems Using Quadratic Programming,” Association for Computational Linguistics, 2015. Accessed: Jun. 18, 2022. [Online]. Available: http://pan.baidu.com/

[4] S. Roy, T. Vieira, and D. Roth, “Reasoning about Quantities in Natural Language,” *Trans Assoc Comput Linguist*, vol. 3, pp. 1–13, Dec. 2015, doi: 10.1162/tacl_a_00118.

[5] A. Mitra and C. Baral, "Learning To Use Formulas To Solve Simple Arithmetic Problems." Accessed: Jun. 18, 2022. [Online]. Available: http://allenai.org/euclid.html
[6] R. Koncel-Kedziorski, H. Hajishirzi, A. Sabharwal, O. Etzioni, and S. D. Ang, "Parsing Algebraic Word Problems into Equations," *Trans Assoc Comput Linguist*, vol. 3, pp. 585–597, Dec. 2015, doi: 10.1162/tacl_a_00160.
[7] Y. Wang, X. Liu, and S. Shi, "Deep Neural Solver for Math Word Problems." Accessed: Jun. 18, 2022. [Online]. Available: https://aclanthology.org/D17-1088/
[8] L. Wang *et al.*, "MathDQN: Solving Arithmetic Word Problems via Deep Reinforcement Learning." Accessed: Jun. 18, 2022. [Online]. Available: www.aaai.org
[9] R. Socher *et al.*, "*Recursive* Deep Models for Semantic Compositionality Over a Sentiment Treebank," Association for Computational Linguistics. Accessed: Jun. 18, 2022. [Online]. Available: http://nlp.stanford.edu/
[10] R. Socher, C. Chiung, Y. Lin, A. Y. Ng, and C. D. Manning, "Parsing Natural Scenes and Natural Language with *Recursive* Neural Networks," 2011. Accessed: Jun. 18, 2022. [Online]. Available: www.socher.org.
[11] Y. Zhang, H. Zhou, and Z. Li, "Fast and accurate neural CRF constituency parsing," in *IJCAI International Joint Conference on Artificial Intelligence*, 2020, vol. 2021-January, pp. 4046–4053. doi: 10.24963/ijcai.2020/560.
[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality." Accessed: Jun. 18, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.
[14] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-Supervised *Recursive* Autoencoders for Predicting Sentiment Distributions." Accessed: Jun. 18, 2022. [Online]. Available: www.socher.org.
[15] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic Pooling and Unfolding *Recursive* Autoencoders for Paraphrase Detection." Accessed: Jun. 18, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2011/hash/3335881e06d4d23091389226225e17c7-Abstract.html
[16] B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning, "Max-Margin Parsing." Accessed: Jun. 18, 2022. [Online]. Available: https://aclanthology.org/W04-3201.pdf
[17] N. D. Ratliff, J. Andrew Bagnell, and M. A. Zinkevic, "Maximum margin planning," in *ACM International Conference Proceeding Series*, 2006, vol. 148, pp. 729–736. doi: 10.1145/1143844.1143936.
[18] M. O'donnell, "RSTTool 2.4-A Markup Tool for Rhetorical Structure Theory." Accessed: Jun. 18, 2022. [Online]. Available: https://aclanthology.org/W00-1434.pdf