HIJM
HORIZON: Indonesian Journal of Multidisciplinary

p-ISSN: 2987-7776
e-ISSN: 2987-7180
Vol. 4, No. 4, 2026

IDENTIFIKASI KUANTITAS RELEVAN PADA SOAL CERITA MATEMATIKA MENGGUNAKAN MODEL BILSTM DAN CRF

Muhamad Taufiq Hidayat[1], Agung Prasetya[2], Joko Iskandar[3]
[1, 2, 3]Universitas Bhinneka PGRI, Jl. Mayor Sujadi No.7, Tulungagung, Jawa Timur, Indonesia
Email: senitaufiq@gmail.com

Article History

Received: 29-05-2026

Revision: 18-06-2026

Accepted: 23-06-2026

Published: 27-07-2026

Abstract. Mathematical word problems often contain information that is not entirely relevant, causing students to make mistakes in selecting important information. This study aims to identify relevant quantities in Indonesian mathematical word problem texts using a sequence labeling approach. This research applies a quantitative approach through computational experiments. The dataset consists of 700 problems collected from elementary school mathematics e-books and annotated using the BIO scheme with B-REL, I-REL, and O labels. The model used is a combination of BiLSTM and CRF, with an 80:20 split for training and testing data. Evaluation was conducted using precision, recall, F1-score, and accuracy, with a focus on the F1-score of the REL class. The results show that the model achieved an F1-score of 0.9726. The dominant error occurred in the classification of REL as O, indicating that quantity spans were still often truncated. However, no BIO rule violations were found, indicating that label sequence consistency was maintained. These results show that the model still has difficulty recognizing some parts of relevant quantities completely. This study is limited to elementary school arithmetic word problems involving basic operations. It also focuses only on identifying relevant quantities, not on solving the final answers to the problems. Future research may add more varied data and compare the BiLSTM-CRF model with other models.

Keywords: Mathematical Word Problems, Relevant Quantities, Sequence Labeling

Abstrak. Soal cerita matematika sering kali memuat informasi yang tidak semuanya relevan, sehingga siswa sering mengalami kesalahan dalam memilih informasi penting. Penelitian ini bertujuan untuk mengidentifikasi kuantitas relevan dalam teks soal cerita matematika berbahasa Indonesia menggunakan *sequence labeling*. Penelitian ini menggunakan pendekatan kuantitatif melalui eksperimen komputasional. Data penelitian terdiri dari 700 soal dari *e-book* matematika sekolah dasar yang dianotasi menggunakan skema BIO (B-REL, I-REL, dan O). Model yang digunakan adalah BiLSTM dan CRF dengan pembagian data latih dan uji sebesar 80:20. Evaluasi dilakukan menggunakan *precision*, *recall*, *F1-score*, dan *accuracy* dengan fokus pada *F1-score* kelas REL. Hasil menunjukkan bahwa model mencapai *F1-score* sebesar 0,9726. Kesalahan dominan terjadi pada klasifikasi REL ke O, yang menunjukkan bahwa *span* kuantitas masih sering terpotong. Namun, tidak ditemukan pelanggaran aturan BIO sehingga konsistensi label tetap terjaga. Hasil tersebut menunjukkan bahwa model masih belum mengenali beberapa bagian kuantitas relevan secara lengkap. Penelitian ini hanya membahas soal aritmatika sekolah dasar dengan operasi dasar. Penelitian ini juga hanya berfokus pada identifikasi kuantitas relevan, bukan pada penyelesaian akhir soal. Penelitian selanjutnya dapat menambah variasi data dan membandingkan model BiLSTM-CRF dengan model lain.

Kata Kunci: Soal Cerita Matematika, Kuantitas Relevan, Pelabelan Urutan

How to Cite: Hidayat, M. T., Prasetya, A., & Iskandar, J. (2026). Identifikasi Kuantitas Relevan pada Soal Cerita Matematika Menggunakan Model BiLSTM dan CRF. *HORIZON: Indonesian Journal of Multidisciplinary*, 4 (4), 5501-5508. http://doi.org/10.54373/hijm.v4i4.6048

PENDAHULUAN

Soal cerita matematika atau *Math Word* Problem (MWP) adalah soal matematika dalam bentuk teks naratif. Teks ini berisi situasi cerita dan memuat operasi hitung. MWP ditulis dengan bahasa alami dan mendekati cara manusia menyampaikan masalah sehari-hari. Dalam proses penyelesaiannya, soal cerita matematika tidak hanya menuntut kemampuan berhitung, tetapi juga kemampuan memilih informasi yang relevan. Tantangan ini juga muncul pada sistem otomatis penyelesaian *Mathematical Word Problems* (MWP). Sistem perlu memetakan bahasa alami menjadi operasi matematika dengan memahami hubungan antara angka, objek, dan tindakan dalam teks (Prasetya, 2024; Vessonen et al., 2024; Yu et al., 2023). Selain itu, sistem harus mampu mengabaikan informasi yang bersifat pengalih sebelum membentuk ekspresi solusi. Tanpa proses pemilahan yang tepat, kesalahan dapat terjadi sejak tahap awal dan memengaruhi hasil akhir. Dalam konteks penelitian ini, kuantitas relevan didefinisikan sebagai nilai numerik beserta satuannya atau objek terkait yang secara langsung digunakan dalam proses perhitungan untuk menjawab pertanyaan pada soal cerita. Kuantitas ini berbeda dari informasi numerik lain yang hanya berfungsi sebagai konteks atau pengalih dan tidak berkontribusi terhadap solusi akhir. Sebagian besar penelitian MWP berfokus pada pembentukan persamaan atau struktur ekspresi solusi, termasuk penggunaan representasi perantara dan strategi optimasi (Huang et al., 2018; Mandal & Naskar, 2021; Wang et al., 2018). Namun, proses identifikasi kuantitas relevan sering kali tidak dieksplisitkan pada level token, sehingga analisis kesalahan menjadi terbatas.

Penelitian ini menerapkan identifikasi kuantitas relevan sebagai tugas *sequence labeling* pada tingkat *token* menggunakan skema BIO dengan label B-REL, I-REL, dan O. Pendekatan ini memungkinkan pelacakan langsung terhadap token yang berkontribusi dalam penyelesaian soal. Model yang digunakan adalah *Bidirectional Long Short-Term Memory* dengan *Conditional Random Fields* (BiLSTM-CRF), yang dikenal efektif dalam menangkap konteks sekuens dan menjaga konsistensi antar label (Huang et al., 2015; Lample et al., 2016; Ma & Hovy, 2016; Sutton & Mccallum, 2010). *Dataset* disusun dari soal cerita aritmetika sekolah dasar berbahasa Indonesia dan direpresentasikan dalam format CoNLL untuk mendukung proses pelatihan dan evaluasi. Kondisi tersebut menunjukkan perlunya pendekatan yang dapat menandai kuantitas relevan secara langsung pada teks soal cerita. Berdasarkan uraian tersebut, penelitian ini diarahkan untuk membentuk *dataset* soal cerita matematika berbahasa Indonesia yang dilabeli menggunakan skema BIO, merancang model BiLSTM-CRF untuk mengenali token kuantitas relevan, mengimplementasikan dan melatih model pada *dataset* yang telah disiapkan, serta menguji performa model menggunakan metrik evaluasi yang sesuai.

Evaluasi dilakukan pada level token menggunakan metrik *precision, recall, F1-score*, dan *accuracy* (Grandini et al., 2020). Fokus utama pengukuran adalah *F1-score* pada kelas REL, yang menggabungkan label B-REL dan I-REL. Pendekatan ini dipilih karena distribusi label O cenderung dominan dalam tugas *sequence labeling*. Selain itu, penelitian ini menganalisis pola kesalahan model, khususnya kesalahan bertipe REL→O yang menunjukkan terjadinya pemotongan span kuantitas pada token lanjutan. Analisis ini memberikan pemahaman lebih dalam terhadap karakteristik kesalahan model dalam mengidentifikasi kuantitas relevan.

METODE

Penelitian ini menggunakan pendekatan *sequence labeling* untuk mengidentifikasi kuantitas relevan pada soal cerita matematika berbahasa Indonesia. *Dataset* disusun dari 700 soal cerita aritmetika sekolah dasar yang dikumpulkan dari *e-book* matematika. Soal dipilih berdasarkan kriteria memiliki narasi, nilai numerik, dan satu pertanyaan dengan jawaban numerik tunggal. Setiap soal kemudian dianotasi menggunakan skema BIO dengan label B-REL, I-REL, dan O untuk menandai posisi kuantitas relevan pada tingkat token, di mana B-REL menunjukkan token awal kuantitas relevan, I-REL menunjukkan token lanjutan dalam satu span kuantitas, dan O menunjukkan token di luar kuantitas relevan.

Teks soal diproses melalui tahap tokenisasi dan normalisasi. Token angka dipetakan menjadi <NUM> dan seluruh huruf diubah menjadi huruf kecil. *Dataset* disimpan dalam format CoNLL yang merepresentasikan pasangan token dan label dalam bentuk sekuens terstruktur. Data kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20.

Model yang digunakan adalah *Bidirectional Long Short-Term Memory with Conditional Random Fields* (BiLSTM-CRF). Lapisan BiLSTM digunakan untuk menangkap konteks dua arah dari urutan token, sedangkan lapisan CRF digunakan untuk memodelkan ketergantungan antar label sehingga menghasilkan urutan label yang konsisten (Huang et al., 2015; Lample et al., 2016; Ma & Hovy, 2016; Sutton & Mccallum, 2010). Representasi token dibentuk melalui *embedding*, kemudian diproses oleh BiLSTM, dilanjutkan dengan lapisan linear untuk menghasilkan skor emisi, dan diakhiri dengan CRF untuk menentukan urutan label optimal.

Pelatihan model dilakukan menggunakan *optimizer Adam* dengan fungsi *loss* berupa *negative log-likelihood* dari CRF (Kingma & Ba, 2015). Proses pelatihan dilakukan dalam beberapa epoch dengan teknik mini-batch serta penerapan *padding* dan *masking* untuk menangani perbedaan panjang *sekuens* (Lample et al., 2016; Ma & Hovy, 2016).

Evaluasi dilakukan pada level token menggunakan metrik *precision, recall, F1-score*, dan *accuracy* (Grandini et al., 2020). Fokus utama pengukuran adalah *F1-score* pada kelas REL yang menggabungkan label B-REL dan I-REL. Pendekatan ini dipilih karena distribusi label O yang dominan dalam tugas *sequence labeling*. Selain itu, analisis kesalahan dilakukan untuk mengidentifikasi pola kesalahan model, khususnya kesalahan klasifikasi REL menjadi O yang berkaitan dengan pemotongan span kuantitas.

HASIL DAN DISKUSI

Pengujian dilakukan pada data uji dari pembagian 80:20 dengan total 140 soal. Evaluasi dilakukan pada tingkat *token* menggunakan metrik *precision*, *recall*, dan *F1-score* untuk setiap label BIO, yaitu B-REL, I-REL, dan O. Selain itu, dihitung pula *F1-score* kelas REL yang merupakan gabungan label B-REL dan I-REL sebagai ukuran utama kinerja model dalam mengidentifikasi kuantitas relevan. Penggunaan *F1-score* kelas REL dipilih karena tujuan utama penelitian ini bukan hanya memperoleh akurasi tinggi, tetapi memastikan bahwa kuantitas yang benar-benar digunakan dalam penyelesaian soal dapat dikenali secara tepat.

Tabel 1. Hasil evaluasi model BiLSTM-CRF

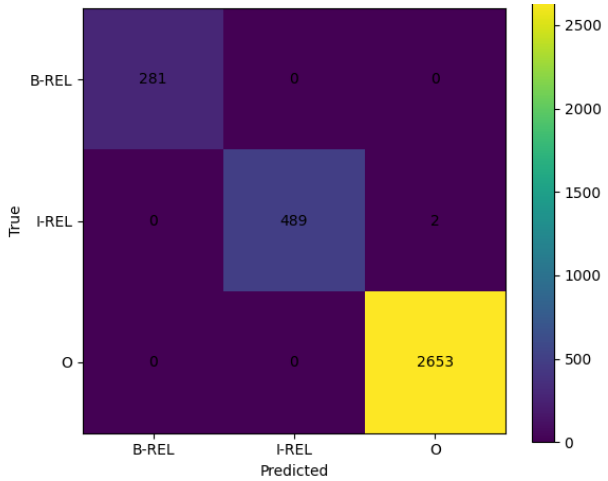
Label	Precision	Recall	F1-Score
B-REL	1.00	1.00	1.00
I-REL	0.98	0.94	0.96
O	0.99	1.00	0.99
Accuracy			0.99
Macro avg	0.99	0.98	0.98
Weighted	0.99	0.99	0.99

Berdasarkan Tabel 1, model BiLSTM-CRF memperoleh *accuracy* sebesar 0,99 dan *F1-score* kelas REL sebesar 0,9726. Hasil ini menunjukkan bahwa model mampu mengidentifikasi kuantitas relevan dengan sangat baik pada soal cerita matematika berbahasa Indonesia. Kinerja terbaik terlihat pada label B-REL yang memperoleh nilai *precision*, *recall*, dan *F1-score* sebesar 1.00. Temuan ini menunjukkan bahwa model mampu mengenali token awal kuantitas relevan secara konsisten. Kondisi tersebut kemungkinan dipengaruhi oleh pola data yang relatif jelas, yaitu token awal kuantitas relevan umumnya berupa angka yang diikuti oleh satuan atau objek tertentu.

Label I-REL memperoleh *precision* sebesar 0.98, *recall* sebesar 0.94, dan *F1-score* sebesar 0.96. Nilai ini masih tergolong tinggi, tetapi lebih rendah dibandingkan label B-REL. Perbedaan tersebut menunjukkan bahwa model lebih mudah mengenali awal kuantitas dibandingkan mempertahankan seluruh bagian kuantitas sampai akhir *span*. Dengan kata lain, model sudah mampu menemukan kuantitas relevan, tetapi masih mengalami kesalahan kecil

ketika kuantitas relevan berbentuk frasa yang lebih panjang. Hal ini terlihat pada beberapa token lanjutan seperti “mempunyai,” “kenarinya,” “Fauzan,” “komplek,” dan “melati” yang seharusnya berada dalam rentang kuantitas relevan, tetapi diprediksi sebagai O.

Label O menunjukkan performa tinggi dengan *recall* sebesar 1.00 dan *F1-score* sebesar 0.99. Hasil ini menunjukkan bahwa model sangat baik dalam mengenali token yang tidak termasuk kuantitas relevan. Kesalahan tipe O→REL hanya berjumlah 1 token, yaitu pada token “lebihnya.” Kondisi ini menunjukkan bahwa model tidak terlalu agresif dalam menandai token nonrelevan sebagai relevan. Dengan demikian, model lebih cenderung berhati-hati dalam menentukan token relevan daripada memperluas prediksi ke token yang tidak diperlukan.



Gambar 1. *Confusion matrix* BiLSTM-CRF

Hasil *confusion matrix* menunjukkan bahwa sebagian besar prediksi berada pada diagonal utama. Hal ini menandakan bahwa model mampu mengklasifikasikan setiap label secara tepat. Label B-REL terdeteksi tanpa kesalahan, sedangkan kesalahan kecil terjadi pada label I-REL yang sebagian diprediksi sebagai O. Pola kesalahan ini memperkuat temuan bahwa tantangan utama model bukan terletak pada pengenalan awal kuantitas, tetapi pada penjagaan kesinambungan token dalam satu *span* kuantitas. Nilai BIO *violation* sebesar 0 juga menunjukkan bahwa urutan label yang dihasilkan tetap konsisten dan mengikuti aturan BIO. Hal ini mendukung peran lapisan CRF dalam menjaga keteraturan transisi label.

Temuan penelitian ini selaras dengan Huang et al., (2015) yang menunjukkan bahwa kombinasi BiLSTM dan CRF efektif untuk tugas penandaan urutan. BiLSTM berperan dalam membaca konteks dari dua arah, sedangkan CRF membantu memilih urutan label yang paling konsisten. Pada penelitian ini, kemampuan tersebut terlihat dari nilai *accuracy* dan *F1-score* yang tinggi, serta tidak ditemukannya pelanggaran aturan sesuai digunakan pada tugas *sequence labeling*, khususnya untuk mengidentifikasi kuantitas relevan pada soal cerita matematika.

Hasil penelitian ini juga sejalan dengan penelitian Huang et al., (2018); Koncel-kedziorski & Hajishirzi (2015); Mandal & Naskar (2021); Wang et al., (2018) yang menempatkan pemilihan kuantitas sebagai tahap penting dalam penyelesaian *Math Word Problem* (MWP). Koncel-kedziorski & Hajishirzi (2015) menempatkan kuantitas pada bagian daun dalam pohon ekspresi. Huang et al., (2018) menggunakan representasi perantara untuk membantu pemetaan teks soal ke bentuk persamaan. Mandal & Naskar (2021) menggunakan klasifikator relevansi kuantitas untuk membedakan kuantitas relevan dan tidak relevan. Wang et al., (2018) memperkenalkan *quantity schema* untuk membantu sistem memilih kuantitas yang sesuai sebelum membentuk ekspresi solusi. Hasil penelitian ini mendukung pandangan tersebut karena kuantitas relevan terbukti dapat diidentifikasi secara otomatis dengan nilai *F1-score* kelas REL sebesar 0.9726.

Meskipun selaras dengan penelitian terdahulu, penelitian ini memiliki fokus yang berbeda. Penelitian terdahulu umumnya menempatkan identifikasi kuantitas sebagai bagian dari proses pembentukan persamaan atau pohon ekspresi. Pada penelitian ini, identifikasi kuantitas relevan diformalkan secara langsung sebagai tugas *sequence labeling* pada tingkat *token*. Dengan pendekatan ini, posisi token yang berkontribusi terhadap penyelesaian soal dapat dilacak secara lebih rinci. Perbedaan ini memberi keuntungan dalam analisis kesalahan karena model tidak hanya dinilai dari hasil akhir, tetapi juga dari ketepatan dalam mengenali batas kuantitas relevan di dalam teks.

Tabel 2. Contoh tabel kesalahan label

Token	Label Benar	Prediksi	Analisis Kesalahan
Mempunyai	I-REL	O	REL→O
Kenarinya	I-REL	O	REL→O
Sudah	I-REL	O	REL→O
Berkicau	I-REL	O	REL→O
Fauzan	I-REL	O	REL→O
Membagi	I-REL	O	REL→O
Kelerengnya	I-REL	O	REL→O
Komplek	I-REL	O	REL→O
Melati	I-REL	O	REL→O
Lebihnya	O	I-REL	O→REL

Analisis kesalahan menunjukkan bahwa kesalahan utama berada pada tipe REL→O sebanyak 9 token, sedangkan kesalahan O→REL hanya 1 token. Temuan ini menunjukkan bahwa model lebih sering gagal mengambil seluruh bagian kuantitas relevan daripada salah menandai token nonrelevan sebagai relevan. Kesalahan REL→O terutama terjadi pada label I-REL. Pola ini menunjukkan bahwa sebagian kuantitas relevan masih terpotong pada token lanjutan. Dampaknya, batas kuantitas yang dikenali model belum selalu utuh. Namun, jumlah

kesalahan yang kecil dan nilai *F1-score* yang tinggi menunjukkan bahwa kelemahan tersebut tidak terlalu besar terhadap kinerja keseluruhan model. Implikasi dari hasil penelitian ini dapat dilihat dari sisi teoretis dan praktis. Secara teoretis, penelitian ini menambah bukti bahwa model BiLSTM-CRF dapat digunakan secara efektif pada tugas *sequence labeling* untuk teks soal cerita matematika berbahasa Indonesia. Penelitian ini juga memperluas kajian MWP dengan menempatkan identifikasi kuantitas relevan sebagai tahap khusus sebelum proses penyelesaian soal. Secara praktis, hasil penelitian ini dapat menjadi dasar pengembangan sistem cerdas yang membantu mengenali informasi penting dalam soal cerita matematika. Sistem seperti ini dapat mendukung pengembangan aplikasi pembelajaran, sistem pembantu penyelesaian soal, atau komponen awal dalam sistem otomatis penyelesaian MWP.

Dengan demikian, hasil penelitian menunjukkan bahwa model BiLSTM-CRF efektif dalam mengidentifikasi kuantitas relevan pada soal cerita matematika. Model mampu mengenali token awal kuantitas secara konsisten, menjaga urutan label tetap sesuai aturan BIO, dan menghasilkan kesalahan yang relatif kecil. Namun, model masih perlu ditingkatkan pada pengenalan token lanjutan dalam *span* kuantitas relevan. Penelitian selanjutnya dapat menambah variasi data, memperluas bentuk soal, dan membandingkan model BiLSTM-CRF dengan pendekatan lain agar kemampuan identifikasi kuantitas relevan menjadi lebih kuat.

KESIMPULAN

Identifikasi kuantitas relevan pada soal cerita matematika dilakukan dengan pendekatan *sequence labeling* menggunakan skema BIO pada tingkat token. Model BiLSTM-CRF terbukti mampu menangkap konteks dalam urutan teks dan menjaga hubungan antarlabel, sehingga hasil prediksi menjadi lebih konsisten. Hasil evaluasi menunjukkan nilai *F1-score* kelas REL sebesar 0,9726. Nilai tersebut menunjukkan bahwa model memiliki kemampuan sangat baik dalam membedakan kuantitas relevan dan informasi yang tidak relevan.

Performa tinggi pada label B-REL dan O menunjukkan bahwa model mampu mengenali awal kuantitas relevan dan *token* yang tidak relevan dengan baik. Sementara itu, kesalahan pada label I-REL menunjukkan bahwa model masih mengalami kesulitan dalam mengenali bagian lanjutan dari kuantitas relevan secara lengkap. Keterbatasan penelitian ini terletak pada data yang hanya mencakup soal aritmetika tingkat sekolah dasar dengan operasi dasar. Penelitian ini juga hanya berfokus pada identifikasi kuantitas relevan, bukan pada penyelesaian akhir soal cerita matematika. Penelitian selanjutnya disarankan untuk menambah variasi data, memperluas jenis soal, dan menguji model pada soal dengan struktur kalimat yang lebih kompleks. Selain itu, penelitian berikutnya dapat membandingkan model BiLSTM-CRF

dengan model lain, seperti model berbasis *Transformer*, serta menghubungkan hasil identifikasi kuantitas relevan dengan sistem penyelesaian *Mathematical Word Problems* secara otomatis.

REFERENSI

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: An Overview*. 1–17. https://doi.org/10.48550/arXiv.2008.05756

Huang, D., Yao, J., Lin, C., Zhou, Q., & Yin, J. (2018). *Using Intermediate Representations to Solve Math Word Problems*. 419–428. https://doi.org/10.18653/v1/P18-1039

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. https://doi.org/10.48550/arXiv.1508.01991

Kingma, D. P., & Ba, J. L. (2015). *Adam: A Method for Stochastic Optimization*. 1–15. https://doi.org/10.48550/arXiv.1412.6980

Koncel-kedziorski, R., & Hajishirzi, H. (2015). *Parsing Algebraic Word Problems into Equations*. *3*, 585–597. https://doi.org/10.1162/tacl_a_00160

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). *Neural Architectures for Named Entity Recognition*. 260–270. https://doi.org/10.18653/v1/N16-1030

Ma, X., & Hovy, E. (2016). *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*. 1064–1074. https://doi.org/10.18653/v1/P16-1101

Mandal, S., & Naskar, S. K. (2021). Classifying and Solving Arithmetic Math Word Problems - An Intelligent Math Solver. *IEEE Transactions on Learning Technologies*, *14*(1), 28–41. https://doi.org/10.1109/TLT.2021.3057805

Prasetya, A. (2024). Identifying Arithmetic Operation in Math Word Problem Based on Recursive Neural Network and Support Vector Machine 1. *JoEICT (Journal of Education and ICT)*, *8*(2), 45–49. https://doi.org/10.29100/joeict.v8i2.7421

Sutton, C., & Mccallum, A. (2010). *An Introduction to Conditional Random Fields*. https://doi.org/10.48550/arXiv.1011.4088

Vessonen, T., Dahlberg, M., Hellstrand, H., Widlund, A., Korhonen, J., Aunio, P., & Laine, A. (2024). Task Characteristics Associated with Mathematical Word Problem - Solving Performance Among Elementary School - Aged Children: A Systematic Review and Meta - Analysis. In *Educational Psychology Review* (Vol. 36, Issue 4). Springer US. https://doi.org/10.1007/s10648-024-09954-2

Wang, L., Zhang, D., Gao, L., Song, J., Guo, L., & Shen, H. T. (2018). *MathDQN: Solving Arithmetic Word Problems via Deep Reinforcement Learning*. 5545–5552. https://doi.org/10.1609/aaai.v32i1.11981

Yu, X., Lyu, X., Peng, R., & Shen, J. (2023). Solving Arithmetic Word Problems by Synergizing Syntax-Semantics Extractor for Explicit Relations and Neural Network Miner for Implicit Relations. *Complex & Intelligent Systems*, *9*(1), 697–717. https://doi.org/10.1007/s40747-022-00828-0