Toxic Detection pada Game: Klasifikasi Intent dan Slot Filling pada Dataset CONDA

Bagus Cipta Pratama, Kosmas Rio Legowo, Muhammad Akmal Fauzan, Rafid Nur Huda
Department of Computer Sciences and Electronics
Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Pendahuluan

Seiring dengan meningkatnya popularitas game online multipemain *toxic behaviour* menjadi masalah serius dalam industri game. Karakteristik ujaran toxic pada chat in-game memiliki perbedaan mencolok dengan platform lain seperti media sosial atau portal berita. Pesan-pesan yang muncul cenderung sangat pendek karena pemain harus mengetik sambil bermain, sementara ujaran yang lebih panjang biasanya hanya muncul pada fase sebelum atau sesudah pertandingan. Dengan sifat pesan dalam game yang seperti ini, pemahaman pada level slot/token kata (Weld et al., 2021b) menjadi sangat penting untuk mendeteksi bahasa toxic di dalam game.

Shared task ini bertujuan untuk memahami karakteristik toksisitas dalam pesan-pesan pada game online. Karena percakapan dalam game online menyerupai percakapan bahasa alami, dibutuhkan pendekatan Natural Language Understanding (NLU) yang mengekstraksi struktur semantik melalui analisis intent dan slot. Dalam kerangka ini, deteksi intent didefinisikan sebagai klasifikasi tujuan atau maksud dari suatu ujaran (atau kalimat), sedangkan slot filling adalah pelabelan setiap kata dalam ujaran sesuai jenis informasi semantik yang dibawanya. Sejumlah literatur menunjukkan bahwa kedua tugas tersebut lebih efektif ketika dilatih secara bersama-sama, karena model dapat memanfaatkan hubungan saling melengkapi antara level intent dan level token; pelatihan bersama ini terbukti menghasilkan kinerja yang lebih baik dibandingkan ketika intent dan slot dipelajari secara terpisah (Zhang et al., 2018). Banyak pendekatan berbasis arsitektur *sequence-to-sequence* dengan RNN dan turunannya (LSTM/GRU) yang dikembangkan untuk memodelkan kedua tugas ini secara bersama, sehingga representasi konteks yang dipelajari untuk slot juga dapat menginformasikan prediksi intent, dan sebaliknya. Pendekatan-pendekatan ini menjadi landasan bagi model yang akan diadaptasi dalam penelitian ini, yaitu Slot-Gated Modeling (Goo et al., 2018) untuk deteksi bahasa toksik *in-game* berbasis intent dan slot.

2 Literature Review

Sebelum *deep learning*, *slot filling* umumnya dimodelkan sebagai *sequence labeling* (misalnya CRF), sedangkan *intent detection* memakai klasifikator terpisah. Pendekatan *pipeline* ini mengabaikan keterkaitan kuat antara intent dan slot, sehingga mendorong munculnya model *joint* berbasis RNN/LSTM.

(Liu and Lane, 2016) mengusulkan model *joint* berbasis RNN/BiLSTM dengan mekanisme *attention* untuk memprediksi intent dan slot secara bersamaan. Arsitektur ini menunjukkan bahwa pelatihan bersama dua task dalam satu jaringan meningkatkan akurasi intent dan F1 slot dibanding pelatihan terpisah, dan menjadi salah satu baseline awal untuk SLU berbasis *neural*.

Penelitian oleh (Wang et al., 2018) juga memperkenalkan Bi-model BLSTM yang memakai dua jaringan BLSTM terpisah untuk intent dan slot, namun saling berbagi representasi tersembunyi. Pertukaran informasi ini terbukti memberi peningkatan kinerja di atas model *attention*-based sebelumnya, dan menegaskan bahwa eksploitasi hubungan intent–slot penting bagi pemahaman semantik.

(Goo et al., 2018) kemudian mengusulkan *Slot-*

Gated Modeling yang secara eksplisit memodelkan interaksi antara intent dan slot melalui mekanisme *gate* di atas BLSTM dengan *attention*. Model ini menghitung representasi intent tingkat kalimat dan representasi slot per token, lalu menggabungkannya lewat *slot gate* untuk mengontrol seberapa besar informasi intent memengaruhi prediksi slot.

Optimasi dilakukan dengan memaksimalkan probabilitas bersama label intent dan urutan slot, sehingga model benar-benar belajar keterkaitan keduanya. Pada dataset ATIS dan Snips, Slot-Gated Modeling mencapai peningkatan konsisten pada akurasi intent, F1 slot, maupun *sentence-level semantic frame accuracy* dibanding Attention BiRNN (Goo et al., 2018). Hal ini menunjukkan bahwa pemodelan eksplisit hubungan intent–slot melalui *gating* lebih efektif daripada sekadar berbagi *attention* atau *hidden state*.

3 Dataset

Dataset yang digunakan adalah CONDA (CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection) , yaitu korpus chat *in-game* dari gim *Dota 2* yang dirancang untuk pemodelan toksisitas berbasis intent dan slot (Weld et al., 2021a). Dataset ini berisi 44,869 ujaran dari 12,152 percakapan dalam 1,921 pertandingan, dengan rata-rata panjang ujaran yang sangat pendek (sekitar 3–6 token), mencerminkan sifat komunikasi cepat saat bermain.

Setiap contoh data terdiri dari teks ujaran, daftar token yang telah diproses, label slot untuk tiap token, serta label intent pada level ujaran. CONDA juga menyediakan metadata konteks seperti ID pertandingan, ID percakapan, dan waktu chat.

3.1 Label Intent (Level Ujaran)

CONDA menggunakan empat kelas intent yang merepresentasikan fungsi ujaran sekaligus tingkat toksisitasnya:

- **Explicit Toxicity (E)**: ujaran dengan kata-kata kasar atau hinaan langsung.
- **Implicit Toxicity (I)**: ujaran bernada sarkastik atau merendahkan tetapi tidak eksplisit.
- **Action (A)**: ujaran yang berkaitan dengan tindakan atau koordinasi di dalam permainan.
- **Other (O)**: ujaran umum yang tidak bersifat toksik.

Distribusinya didominasi oleh kelas O (73.9%), sedangkan toksisitas eksplisit dan implisit mencakup sekitar 19.7% dari seluruh data.

3.2 Label Slot (Level Token)

Setiap token dalam ujaran diberi label salah satu dari enam kategori:

- **T** (Toxicity): kata bernada ofensif.
- **C** (Character): nama atau singkatan hero dalam game.
- **D** (Dota-specific): istilah teknis dalam game.
- **S** (Slang): istilah slang khas komunitas.
- **P** (Pronoun): pronomina.
- **O** (Other): token lain yang tidak termasuk kategori sebelumnya.

Mayoritas token berada pada kategori O (63.6%), diikuti P (13.5%) dan S (11.2%).

4 Tantangan

Dataset CONDA memiliki distribusi kelas yang sangat tidak seimbang, di mana label 'Other' (O) mendominasi secara signifikan dibandingkan label target seperti 'Toxic', 'Dota', dan yang lainnya. Hal ini menyulitkan model untuk menangkap pola pada kelas minoritas tanpa bantuan teknik reweighting loss atau augmentasi.

Kemudian, penerapan teknik Data Augmentaation berbasis sinonim pada code kedua menghadapi tantangan 'Semantic Drift'. Karena dataset bersifat domain-spesific (Game Slang), penggunaan thesaurus umum seperti WordNet sering kali menghasilkan kalimat sintetik yang tidak koheren dengan konteks permainan sehingga sedikit menurunkan performa model pada data validasi asli.

Terakhir, tantangan utama dalam Joint Learning adalah menyeimbangkan optimasi antara Intent Classification dan Slot Filling. Dengan metrik Joint Semantic Accuracy (JSA) yang ketat, kesalahan prediksi pada satu token slot saja akan menggugurkan prediksi intent yang sudah benar, sehingga model memerlukan konvergensi yang presisi di kedua task secara simultan.

5 Hasil dan Pembahasan

Pada bagian ini dipaparkan kinerja model yang diusulkan serta analisis trainingnya, dengan menyoroti perbedaan antara konfigurasi baseline dan

model yang telah dioptimasi. Selain itu, juga akan dibahas analisis terkait terjadinya penurunan metrik tertentu pada model yang sudah dioptimasi.

5.1 Kinerja Model

Untuk mengevaluasi efektivitas arsitektur Slot-Gated Joint NLU dalam menangani tugas klasifikasi intent dan pengisian slot (ner) pada dataset CONDA, kami membandingkan dua skenario eksperimen utama:

1. Eksperimen 1 (Baseline): Menggunakan model dasar Slot-Gated BiLSTM dengan inisialisasi embedding acak, tanpa penanganan ketidakseimbangan kelas, tanpa augmentasi data, dan tanpa regularisasi weight decay.

2. Eksperimen 2 (Optimized): Menggunakan arsitektur yang sama, tetapi ditambah dengan empat strategi optimasi, yaitu:

   (a) Pre-trained Embeddings: Inisialisasi menggunakan Word2Vec yang dilatih pada domain-spesific
   (b) Data Augmentation: Augmentasi berbasis sinonim (WordNet) untuk memperkaya variasi data latih
   (c) Weighted Loss: Penerapan Class Weighting pada fungsi Cross Entropy untuk menangani ketidakseimbangan kelas
   (d) Regularization: Penambahan Weight Decay (1e-4) pada optimizer untuk mencegah overfitting

Tabel 1 memperlihatkan rangkuman perbandingan performa kedua model

Table 1: Perbandingan Eksperimen 1 dan Eksperimen 2

Metric	Eksperimen 1	Eksperimen 2
JSA	0.8726	0.8523
F1(E)	0.83	0.83
F1(I)	0.72	0.65
F1(T)	0.96	0.96
F1(D)	0.93	0.94
F1(S)	0.99	0.99

Berdasarkan tabel tersebut, terlihat adanya karakterisitk kinerja yang kontras. Model Baseline menunjukkan performa yang lebih seimbang secara keseluruhan, menghasilkan JSA tertinggi sebesar 86.95%. Sebaliknya, Model Eksperimen 2 mengalami penurunan pada JSA menjadi 85.31%, tetapi mencatat peningkatan konsisten pada seluruh metrik Slot Filling (T, D, dan S)

5.2 Analisis Grafik Loss dan Overfitting

Perbedaan perilaku pembelajaran kedua model terlihat jelas melalui grafik Loss selama proses pelatihan

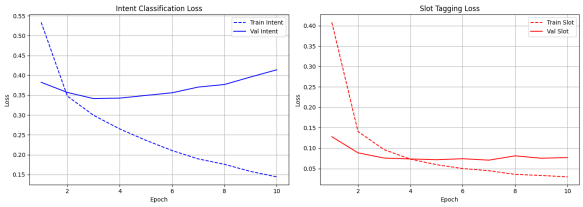

Figure 1: Kurva Loss Eksperimen 1

Pada model baseline, terlihat fenomena overfitting pada task Intent Classification. Kurva loss intent pada gambar 1 menunjukkan validation loss stagnan dan cenderung meningkat, sementara loss train terus turun mendekati nol. Hal ini mengindikasikan bahwa model terlalu menghafal data latih, tetapi model kesulitan menggeneralisasi intent pada data validasi yang belum pernah dilihat. Sementara itu, pada kurva loss slot tagging, jarak antara kurva train dan validation untuk slot relatif kecil dan tidak menunjukkan kenaikan drastis, sehingga kemampuan generalisasi untuk slot tagging masih cukup baik, meskipun model mulai overfit pada sisi intent.

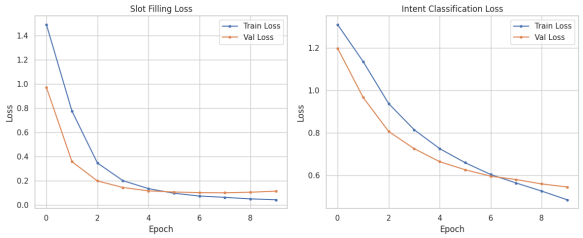

Figure 2: Kurva Loss Eksperimen 2

Sebaliknya, pada eksperimen kedua yang kami lakukan beberapa optimasi, terlihat pada gambar 2 bahwa kurva intent dan slot cenderung stabil melandai pada akhir epoch. Celah antara Train Loss dan Validation Loss relatif kecil, menandakan model yang dilakukan optimasi memiliki kemampuan generalisasi yang cukup baik pada tugas klasifikasi ini.

Limitations

This document does not cover the content requirements for ACL or any other specific venue. Check the author instructions for information on maxi-

mum page lengths, the required "Limitations" section, and so on.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Chih-Wen Goo, Guang-Lai Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung (Vivian) Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *North American Chapter of the Association for Computational Linguistics*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *ArXiv*, abs/1609.01454.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. In *North American Chapter of the Association for Computational Linguistics*.

Henry Weld, Guanghao Huang, Jean Lee, T. Zhang, Kunze Wang, Xinghong Guo, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2021a. Conda: a contextual dual-annotated dataset for in-game toxicity understanding and detection. *ArXiv*, abs/2106.06213.

Henry Weld, Xiaoqi Huang, Siqi Long, Josiah Poon, and Soyeon Caren Han. 2021b. A survey of joint intent detection and slot-filling models in natural language understanding. *CoRR*, abs/2101.08091.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S. Yu. 2018. Joint slot filling and intent detection via capsule neural networks. *ArXiv*, abs/1812.09471.

A Example Appendix

This is an appendix.