

Reference-Free Consensus Evaluation of Machine Translation: Ensembling LLM Error-Span Judgments with Neutral Cross-Verification and STAPLE Reliability Fusion

Assaf Siani

assaf@lexicala.com

Independent Submission, WMT 2026 Shared Task on Automated Translation Quality Evaluation

Abstract

We present a reference-free system for the WMT 2026 automated translation-quality-evaluation shared task, covering all three subtasks: character-level error-span detection with Minor/Major severity, segment-level quality scoring on a 0–100 scale, and binary error-free classification. Rather than treating a single large language model (LLM) as an oracle, the system treats LLMs as fallible annotators. Three heterogeneous judges independently propose target-side error spans; a neutral cross-verification round converts these open-ended proposals into a complete, anonymized candidate-by-judge vote matrix; a character-level adaptation of STAPLE estimates latent error truth and per-judge reliability without references; and a two-state HMM imposes minimal span structure. Omissions are handled by a dedicated source-coverage pass rather than by scanning the target. Calibrating the single free prior against the official metric with human gold yields several findings: plain majority voting matches or beats STAPLE; the error-rate prior barely matters once outputs are schema-conformant; precision—judge over-flagging at roughly $3\times$ the human rate—is the binding constraint; and LLM-derived silver gold cannot calibrate the prior without circularity. The system was deployed across seven language pairs with partial-failure-tolerant, resumable execution.

Keywords: MT evaluation; error-span annotation; quality estimation; LLM-as-a-judge; label fusion; STAPLE; consensus.

1 Introduction

Machine-translation (MT) evaluation increasingly requires more than a scalar system ranking. Error-Span Annotation (ESA) offers localized, severity-aware feedback while retaining a holistic score, occupying a useful middle ground between coarse direct assessment and expert-intensive MQM annotation (Kocmi et al., 2024). The WMT 2026 setting asks for three compatible views of the same segment: character error spans with severity, a 0–100 quality score, and an error-free decision. Crucially, inference is reference-free.

Reference-free inference with LLM judges is deceptively hard. A single judge can be fluent yet lenient, overconfident, prompt-sensitive, and inconsistent across severity scales. Multiple judges improve coverage but create a second problem: they propose different, overlapping intervals, and silence on an interval does not reveal whether a judge rejected it or never considered it. Naive union maximizes recall while inheriting every false positive; naive intersection is too conservative.

We resolve this by converting open-ended proposals into explicit shared decisions and then applying probabilistic label fusion. Our contributions are: (i) a two-round protocol that preserves independence during proposal generation and builds a complete, anonymized candidate-by-judge vote matrix during cross-

verification; (ii) a character-level adaptation of STAPLE with MAP regularization and an HMM decoder that fuses overlapping intervals while estimating judge reliability without references; (iii) a separate source-coverage mechanism for omissions; (iv) a severity and quality layer grounded in local impact, official severity weights, and inverse-variance fusion; and (v) an empirical account, calibrated against the official metric with human gold, of what the consensus machinery does and does not buy.

2 Related Work

String- and embedding-based metrics such as BLEU (Papineni et al., 2002), chrF (Popović, 2015), and COMET (Rei et al., 2020) give useful external diagnostics but do not yield localized, severity-aware spans. LLM-as-a-judge evaluation scales naturally yet inherits systematic bias: position effects, superficial-quality preferences, leniency, and vulnerability to framing (Wang et al., 2024; Zhou et al., 2024; Shi et al., 2025). We observe an analogous label-framing effect—describing pooled candidates as ‘flagged errors’ materially increases acceptance—and therefore verify candidates neutrally. STAPLE (Warfield et al., 2004) estimates a probabilistic latent segmentation while learning each rater’s sensitivity and specificity; its rater-weighting idea transfers well to target-side character masks and, unlike simple voting, lets an explicit clean judgment count as negative evidence.

The WMT Quality Estimation shared tasks have progressively moved toward fine-grained, reference-free evaluation, and recent findings ask whether LLMs are closing the gap to human quality estimation (Zerva et al., 2024). Our system is squarely in that line: reference-free at inference and LLM-based throughout, but departing from the single-judge paradigm by making disagreement explicit and fusing it probabilistically. The MQM framework (Lommel et al., 2013) supplies the severity vocabulary our Minor/Major labels inherit.

3 Task Formulation

For each source segment x and system hypothesis y of N Unicode characters after NFC normalization, the system predicts a set of half-open target intervals $[a, b]$, each labeled Minor or Major; a quality score $Q \in [0, 100]$; and a Boolean error-free decision. Offsets are computed in the exact submitted hypothesis, satisfying $0 \leq a < b \leq N$. Omissions are emitted through the task’s dedicated per-translation omission field rather than as target character spans. Task-1 outputs are scored by the organizers’ MPP metric (Match with Partial overlap and Partial credit, micro-averaged); Task 2 by correlation with human ESA scores; Task 3 by Matthews correlation.

4 System Architecture

The pipeline has three online (API) stages and a deterministic offline tail. (1) Independent annotation: three heterogeneous judges—in the deployed runs Gemini 2.5 Flash, DeepSeek V4 Flash, and Qwen-Plus—each receive the source, the raw translation, language direction, domain, and a diagnostic rubric, and return zero or more target spans with a severity-oriented explanation and a holistic 0–100 score. Calls are independent and parallel; judges never see one another’s proposals. (2) Neutral cross-verification: raw spans are pooled per segment into an anonymized, shuffled candidate list; every judge then votes accept/reject on every candidate and rates the benefit of correction on a bounded scale (1 slight, 2 clear, 3 dramatic), yielding a complete candidate-by-judge matrix. (3) Source-coverage: a dedicated pass detects omissions from the source side. The offline tail performs consensus fusion, quality computation, schema validation, and serialization. Resume-aware JSONL checkpoints let interrupted API runs continue without recomputing completed judgments, and partial judge failures are tolerated: a segment retaining two of three judges is still fused.

5 Consensus Fusion

Character projection. For judge j and character c , $D_{jc} = 1$ when c lies inside any candidate that j accepted, and 0 otherwise. Projection turns partially overlapping

intervals into a common lattice and makes every explicit rejection informative.

STAPLE E-step. Let $T_c \in \{0, 1\}$ be the latent truth and $\gamma = P(T_c = 1)$ the rarity prior. Judge j has sensitivity $p_j = P(D_{jc} = 1 \mid T_c = 1)$ and specificity $s_j = P(D_{jc} = 0 \mid T_c = 0)$. Conditional independence gives the posterior error probability

$$W_c = \frac{\gamma \prod_j (p_j^D (1 - p_j)^{1-D})}{\gamma \prod_j (p_j^D (1 - p_j)^{1-D}) + (1 - \gamma) \prod_j ((1 - s_j)^D s_j^{1-D})}$$

evaluated in log space. Abstaining or failed judges drop out of the product; they lower confidence without counting as negative votes.

MAP M-step. Maximum-likelihood estimates are unstable for short segments and three raters, so Beta-prior pseudo-counts of strength α regularize sensitivity toward $r_j = 0.5$ and specificity toward $s_j = 0.95$:

$$p_j \leftarrow \frac{\sum_c (W_c D_{jc}) + \alpha r_j}{\sum_c (W_c) + \alpha}$$

$$s_j \leftarrow \frac{\sum_c ((1 - W_c)(1 - D_{jc})) + \alpha s_0}{\sum_c (1 - W_c) + \alpha}$$

The asymmetric anchors are deliberate: a high-specificity prior and small γ break the label-swap symmetry and stop the model from painting broad clean regions as errors. EM stops when $\max_c |\Delta W_c| < 10^{-4}$.

HMM smoothing. Independent posteriors contain single-character spikes and holes. A two-state HMM uses the STAPLE likelihoods as emissions with persistence $\rho = P(T_c = 1 \mid T_{c-1} = 1) = 0.85$; to hold the stationary error mass γ , the clean→error transition is $\gamma(1-\rho)/(1-\gamma)$. Viterbi decoding yields the most probable clean/error path, and error runs shorter than two characters are pruned.

6 Severity, Omissions, Error-Free

Reliability-weighted severity. Each fused span inherits impact ratings from the accepting judges whose candidates overlap it. With judge reliability $\rho_j = (p_j + s_j)/2$ and weight $w_j = \max(\log[\rho_j/(1-\rho_j)], \epsilon)$, the fused local impact is $Z_m = \sum_j w_j I_{jm} / \sum_j w_j$ over accepting judges, where $I_{jm} \in \{1, 2, 3\}$ is the impact judge j assigned to span m during cross-verification (1 = slight, 2 = clear, 3 = dramatic). The label is Major when $Z_m \geq 2.0$ and Minor otherwise—an absolute boundary following the prompt semantics (1 = meaning preserved, 2 = clear distortion), chosen over a corpus-relative threshold that would let unrelated batch composition flip labels.

Source-coverage omissions. An omission has no natural target span, so a dedicated pass asks each judge to align source meaning units against the unmodified translation and report uncovered units with impact. An omission is

fused when at least two judges report non-zero severity, becoming Major at fused impact 2.0, and is routed to the dedicated omission field. Treating omission as a source-coverage judgment—not as target-span scanning—avoids a category error that otherwise loses omissions on the most degraded translations.

Error-free classification. The decision is a conjunction: no fused text span, no fused omission, and median holistic adequacy at least 85. The adequacy floor guards against diffuse mistranslation or hallucination that does not localize cleanly. The serializer also exports the organizers’ span-only alternative so both rules coexist transparently.

7 Segment-Level Quality Estimation

Two estimators are fused by inverse variance. Estimator A (debiased holistic): each judge j emits a holistic 0–100 score $h_{j,s}$ on segment s , but judges differ systematically in leniency, so we debias before averaging. For judge j we estimate a global offset b_j —its mean deviation, across all segments it rated, from the ensemble center c_s (the per-segment mean of the judges’ scores, with $|J_s|$ the number of judges scoring segment s):

$$b_j = \text{mean}_s(h_{j,s} - c_s), \quad c_s = \left(\frac{1}{|J_s|}\right) \sum_j (h_{j,s})$$

Subtracting b_j places every judge on a common scale. The segment estimate q_A is a reliability-weighted combination of the offset-corrected scores $h_{j,s} - b_j$, with their dispersion supplying variance v_A ; this keeps the judges’ global fluency/adequacy signal while removing persistent per-judge leniency or harshness. Estimator B (span-derived risk): with official weights 2.5 (Minor) and 10 (Major), a confidence-weighted, length-normalized penalty P maps through a survival link $q_B = 100 \cdot \exp(-P/\beta)$, $\beta = 5000$.

Formally, with fused span m carrying confidence conf_m (its posterior mass under the fusion—STAPLE’s soft posterior, or vote share under majority) and official severity weight $\omega(\text{sev}_m) \in \{2.5, 10\}$, and reference length $L_0 = 500$ characters against the segment’s actual length L :

$$P = \left(\frac{L_0}{L}\right) \sum_m (\text{conf}_m \omega(\text{sev}_m))$$

The L_0 / L term normalizes for segment length, so segments with the same density of errors receive comparable penalty regardless of raw character count; weighting by conf_m lets a low-confidence span contribute partial penalty rather than an all-or-nothing count. Bootstrap resampling of span correctness gives v_B (floored at 25). The fused score is

$$Q = \frac{\frac{q_A}{v_A} + \frac{q_B}{v_B}}{\frac{1}{v_A} + \frac{1}{v_B}}, \quad v_Q = \frac{1}{\frac{1}{v_A} + \frac{1}{v_B}}$$

If holistic ratings are absent, Q degrades gracefully to q_B ; a segment whose every judge fails is marked unscorable and assigned a low, high-variance score rather than a spurious perfect one. Alongside Q we also export the organizers’ mechanical score $\max(0, 100 - 10 \cdot n_{\text{major}} - 2.5 \cdot n_{\text{minor}})$, keeping the distinction between our modeling and the task-prescribed derivation explicit.

8 Experiments and Findings

We integrated the official Task-1 scorer and used human gold from WMT24 ESA and WMT23 MQM to calibrate the single free prior γ . Five findings stand out.

(1) Precision is the binding constraint. On genuine MT output the judges over-flag: their union covers roughly $3\times$ the gold character mass, and per-pair precision sits at 0.13–0.24. No setting of the prior repairs this; reducing upstream over-flagging is where the remaining headroom is.

(2) γ is a decision prior, not the corpus error rate, and it barely matters. On WMT23 he–en the gold density is 10.8% yet the F-optimal γ is 1%, because γ counteracts correlated over-flagging. At full scale the optimal γ varies by pair (0.03–0.15) but $F(\gamma)$ is flat near its maximum: a single global $\gamma = 0.05$ costs at most ≈ 1.2 F-points, so per-pair tuning is low value.

Pair	Precision	Recall	F-opt γ
cs–uk	0.13	0.42	0.05
en–ru	0.15	0.39	0.03
en–is	0.24	0.31	0.15

Table 1: Per-pair Task-1 precision/recall and F-optimal γ against WMT24 ESA gold (500 hypotheses/pair). Precision is uniformly the bottleneck; the F-optimal γ varies but the curves are flat near their maxima.

(3) Schema conformance beats algorithm tuning. Emitting omissions through the dedicated field rather than as character spans raised F by 0.025 on average—twice the entire effect of tuning the prior—because span-encoded omissions had been pure false positives. JSON-lines formatting and Boolean labels were likewise the difference between a valid and a rejected submission.

(4) Majority voting is competitive with STAPLE. In gold ablations, character-level majority vote matched or beat the full STAPLE+HMM pipeline on both tasks (Task-1 F 0.260 vs. 0.246; Task-2 Pearson 0.538 vs. 0.491). STAPLE’s complexity is not clearly earned on these three-rater segments; we therefore submit both variants.

Variant	Task-1 F	Task-2 r
STAPLE + HMM	0.246	0.491

Majority vote	0.260	0.538
---------------	-------	-------

Table 2: Fusion ablation on the three ESA gold pairs. Task-1 F uses the official MPP metric; Task-2 r is Pearson correlation with human ESA scores. Majority voting matches or beats STAPLE on both tasks.

(5) Silver gold cannot calibrate the prior. Deriving gold from a strong LLM to tune γ is circular—it is the very bias γ must counteract—and selects $\gamma = 0.15$ where the human-gold optimum is 0.01, reporting $F = 0.65$ where the honest score is 0.35. For context, our F of 0.20–0.30 is competitive: the best WMT25 auto-rater averaged 13.5% micro- $F1$, and human–human agreement is near 47% (metrics not directly comparable).

(6) A negative result: repair-validation. An earlier design asked each judge to repair a proposed span and re-score the repaired variant under blinded batching, on the theory that a genuine error yields a positive counterfactual quality delta. In practice the filter required two unreliable events—localized repair and calibrated re-scoring. Repairs frequently changed text outside the requested interval, and absolute scores saturated or moved coarsely, so the delta was dominated by noise. We replaced this loop with the direct categorical cross-verification of Section 4, which needs no repair and produces a clean vote matrix. We report the failure because the appealing counterfactual framing is a natural trap.

Diagnostic correlation. As a sanity check, segment scores Q on a 40-segment held-out set correlate positively with reference-based BLEU, chrF, and COMET despite Q using no reference at inference, indicating a shared quality signal. We stress that this is diagnostic only: the sample is small and non-independent, the comparators measure different constructs, and no superiority is claimed.

(7) Task 3 (error-free), MCC. We evaluate the error-free decision against gold derived from the ESA annotations (error-free iff the human ESA score is at least 85 and there are no gold spans). The classes are imbalanced (58% error-free on cs-uk, 17% on en-is), so we report Matthews correlation (MCC). Our adequacy-gated conjunction attains MCC 0.301 (cs-uk) and 0.415 (en-is), beating the organizers’ span-only baseline (0.277 and 0.399) on both pairs: the adequacy floor removes false-positive error-free labels—hallucinated or off-topic translations that produce no localizable span yet are not publishable. Error-free precision/recall are 0.73/0.61 (cs-uk) and 0.44/0.65 (en-is).

Pair	n	gold EF	MCC (ours)	MCC (base)
cs-uk	99	58%	0.301	0.277
en-is	100	17%	0.415	0.399

Table 3: Task-3 error-free MCC against ESA-derived gold. Our adequacy-gated rule beats the span-only baseline on both pairs; the imbalanced classes make MCC the appropriate metric.

(8) Isolating cross-verification from fusion. Neutral cross-verification is separable from the fuser and beneficial on its own: on 178 development candidates, framing pooled candidates neutrally rather than as ‘flagged errors’ raised the number rejected by all three judges from 17 to 43, removing stylistic false positives while preserving spans in heavily degraded probes. Both fusers consume this same cross-verified vote matrix, so the majority-vs-STAPLE comparison (Table 2) isolates only the fusion algorithm layered on top of it. STAPLE’s marginal additions over majority are the per-judge reliability weights (reused in severity and quality), soft character posteriors, and the HMM contiguity prior; on three-rater segments these do not raise F , so we recommend majority vote as the primary submission and keep STAPLE as the second permitted variant.

9 Deployment and Language Pairs

The system was run on seven WMT 2026 language pairs: English–Northern Sami, Czech–German, Chinese–Japanese, English–Arabic (Egyptian and Modern Standard merged), English–Estonian, English–Indonesian, and English–Icelandic (including the GAMBIP challenge set). Direction and domain are routed per record from the item identifier, and language codes are mapped to explicit variety names so judges grade against the correct variety—critical where Egyptian Arabic and MSA are mixed. Execution is resumable and partial-failure tolerant: each stage checkpoints incrementally, and a segment losing one of three judges is still fused on the remaining two, which the source-coverage design shows is sufficient for omission. This heterogeneous-provider ensemble also proved operationally robust to per-provider quota and availability limits during large-scale runs. Table 4 reports the per-pair volume of the submitted output.

Pair	Items	Hyps.	Spans	Omis.	Error-free	EF%
en-se	493	13,311	31,750	5,800	833	6.3
cs-de	1,530	39,198	55,751	10,873	10,333	26.4
zh-ja	690	20,455	34,257	7,700	2,919	14.3
en-ar	1,917	15,782	19,411	3,954	3,936	24.9
en-et	712	21,004	16,978	3,196	4,380	20.9
en-id	331	10,261	10,876	2,831	2,758	26.9
en-is	2,028	24,576	31,887	538	2,725	11.1
Total	7,701	144,587	200,910	34,892	27,884	19.3

Table 4: Per-pair volume of the submitted output (majority variant): source items, scored hypotheses, predicted error spans, omissions routed to the dedicated field, and error-free decisions, with totals across all seven pairs.

10 Discussion

What the consensus machinery buys is epistemic bookkeeping rather than a single accuracy trick.

Independent generation separates proposal from adjudication; neutral cross-verification converts different proposal spaces into shared explicit judgments; STAPLE separates latent truth from annotator performance; the HMM imposes a minimal structural prior; and the source-coverage pass models absence in the source–target relation. No single component proves correctness, but each removes a specific ambiguity that a naive union or intersection would silently resolve the wrong way. That the far simpler majority vote is competitive (Table 2) does not negate this: the disciplined item space and neutral framing are shared by both fusers, and they are where most of the benefit accrues.

The labels reference-free and calibration-free deserve scrutiny. Inference uses no reference translation and rater weights are estimated from agreement, but the prompts encode a quality theory, providers embed training priors, and γ , α , s_0 , ρ , the pruning rule, the adequacy floor, and β are fixed choices partly informed by development data. The honest claim is therefore reference-free inference with prior-encoded design. Finally, STAPLE assumes conditional independence across raters; heterogeneous providers reduce but do not remove correlated errors, so modeling rater correlation is a natural next step.

11 Limitations

Development sets are small and partly synthetic and do not substitute for blind evaluation or native adjudication. Task-1 F-scores remain low, showing that consensus and prompt design do not solve recall and boundary matching. STAPLE’s conditional-independence assumption is strained when judges share training data or instruction-tuning conventions, so the vote matrix can be overconfident when all judges share a misconception, especially for low-resource pairs. Several constants are fixed design choices informed by development data; the strongest defensible claim is reference-free inference with prior-encoded design, not calibration-free evaluation. Proprietary model drift and decoding settings remain external reproducibility constraints; stored JSONL artifacts are treated as the immutable record.

12 Reproducibility

The judges are gemini-2.5-flash, deepseek-v4-flash, and qwen-plus, queried at temperature 0.3 for span annotation and 0 for terse single-score calls, with up to 8000 output tokens. Fixed constants: rarity prior $\gamma = 0.05$; MAP pseudo-count $\alpha = 10$ with sensitivity anchor $r_j = 0.5$ and specificity anchor $s_0 = 0.95$; HMM persistence $\rho = 0.85$ with two-character run pruning; severity cut $Z_m = 2.0$; quality $\beta = 5000$ with reference length $L_0 = 500$; adequacy floor 85. Offsets are half-

open NFC character indices; runs checkpoint incrementally and are deterministic given identical model responses. Because the judges are proprietary and may drift, the stored JSONL judgments are the immutable record against which offline fusion changes should be compared.

13 Conclusion

We build a coherent reference-free WMT evaluation system by treating LLM outputs as noisy annotations to be verified and fused rather than as final truth. Independent generation improves candidate recall; neutral cross-verification creates a common item space and reduces framing-induced acceptance; STAPLE separates latent error from rater performance; the HMM imposes a minimal structural prior; and a source-coverage pass models absence correctly. Calibrated against the official metric with human gold, the analysis is candid about limits: precision-limited judges cap F-score, simple majority voting rivals the probabilistic model, and schema conformance outweighs parameter tuning. These lessons, more than any single component, are the contribution.

References

- Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., Sachan, M., and Shmatova, M. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 1440–1443, Miami, Florida, USA. Association for Computational Linguistics.
- Lommel, A. R., Burchardt, A., and Uszkoreit, H. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer* 35.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Popović, M. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., and Vosoughi, S. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the Joint International Conference on Natural Language Processing and the Asia-Pacific Chapter of the Association for*

- Computational Linguistics (IJCNLP-AACL), pages 292–314. Association for Computational Linguistics.
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Warfield, S. K., Zou, K. H., and Wells, W. M. 2004. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image M. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 1440–1443, Miami, Florida, USA. Association for Computational Linguistics.
- Zerva, C., Blain, F., C. De Souza, J. G., Kanojia, D., Deoghare, S., Guerreiro, N. M., Attanasio, G., Rei, R., Orasan, C., Negri, M., Turchi, M., Chatterjee, R., Bhattacharyya, P., Freitag, M., and Martins, A. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.
- Zhou, H., Huang, H., Long, Y., Xu, B., Zhu, C., Cao, H., Yang, M., and Zhao, T. 2024. Mitigating the bias of large language model evaluation. In *Proceedings of the China National Conference on Computational Linguistics (CCL)*, pages 1310–1319.