From Headlines to Narratives

Mathematical Theory of LLM-Based Narrative Extraction

Prof. Dr. Joerg Osterrieder

Advanced NLP and Machine Learning Theory

September 19, 2025

Mathematical Framework Overview

Scope: Complete mathematical treatment of narrative extraction using Large Language Models. From raw text embeddings through topic discovery to narrative generation. Emphasis on theoretical foundations of NLP methods.

Mathematical Foundations of Text Embeddings

Word2Vec Objectives
Skip-gram objective function:

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

Softmax formulation:

$$p(w_O|w_I) = \frac{\exp(v_{w_O}^T v_{w_I})}{\sum_{w=1}^{W}\exp(v_w^T v_{w_I})} \tag{2}$$

Negative Sampling Approximation:

$$\log\sigma(v_{w_O}^T v_{w_I}) + \sum_{k=1}^{K}\mathbb{E}_{w_k\sim P_n(w)}[\log\sigma(-v_{w_k}^T v_{w_I})] \tag{3}$$

where $P_n(w) = \frac{U(w)^{3/4}}{\sum_w U(w)^{3/4}}$ is the noise distribution.

GloVe and FastText Extensions

GloVe: Co-occurrence Matrix Factorization
Objective function:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{4}$$

FastText: Subword Information

$$s(w, c) = \sum_{g \in G_w} z_g^T v_c \tag{5}$$

where G_w is the set of n-grams for word w.
Character n-gram representations:

- Handle out-of-vocabulary words
- Morphological information preservation
- Shared representations across related words

Key Insight: Subword units capture morphological patterns essential for narrative understanding.

ELMo: Bidirectional LSTM Language Models

$$h_{LM} = [\vec{h}_{LM}; \overleftarrow{h}_{LM}] \tag{6}$$

GPT: Autoregressive Objective

$$L = \sum_i \log P(u_i | u_{i-k}, ..., u_{i-1}; \Theta) \tag{7}$$

BERT: Bidirectional Transformer
MLM objective: $\mathcal{L}_{MLM} = -\mathbb{E}_{\mathcal{D}} \sum_{m \in M} \log P(x_m | x_{\setminus M})$
NSP objective: $\mathcal{L}_{NSP} = -\mathbb{E}_{(S_A, S_B)} \log P(y | [CLS], S_A, [SEP], S_B)$
Key Advancement: Context-dependent representations vs. static embeddings.

Sentence and Document Embeddings

Universal Sentence Encoder (USE)
Deep Averaging Network: $\text{DAN}(x) = \text{DNN}(\frac{1}{n}\sum_{i=1}^{n} x_i)$
Transformer variant: $\text{USE}_T = \text{Transformer}([w_1, ..., w_n])$
Doc2Vec: Paragraph Vector
Distributed Memory (PV-DM):

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in d} \log P(w | w_{\text{context}}, d) \tag{8}$$

Sentence-BERT: Siamese Networks

$$u = \text{BERT}(S_A), \quad v = \text{BERT}(S_B) \tag{9}$$

Classification: $o = \text{softmax}(W_t(u, v, |u - v|))$
Triplet loss: $\mathcal{L} = \max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$

Embedding Space Geometry and Distance Metrics

Manifold Hypothesis

Narratives lie on low-dimensional manifolds: $ID = \lim_{r \to 0} \frac{\log \mathbb{E}[N(r)]}{\log r}$

Distance Metrics for Narrative Similarity:

Cosine similarity: $\cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||}$

Earth Mover's Distance:

$$EMD(P, Q) = \min_{\gamma \in \Pi(P,Q)} \sum_{i,j} \gamma_{ij} ||x_i - y_j|| \tag{10}$$

Optimal Transport Formulation:

$$W_p(P, Q) = \left(\inf_{\gamma \in \Pi(P,Q)} \int ||x - y||^p d\gamma(x, y) \right)^{1/p} \tag{11}$$

Anisotropy Problem: avg-cos $= \frac{2}{n(n-1)} \sum_{i<j} \cos(v_i, v_j)$

Applications: Document similarity, semantic search, narrative clustering.

Topic Modeling and Narrative Discovery

Latent Dirichlet Allocation

Generative Process

For each document d:

1. Draw topic distribution: $\theta_d \sim \text{Dir}(\alpha)$
2. For each word position n:
   - Draw topic: $z_{dn} \sim \text{Multinomial}(\theta_d)$
   - Draw word: $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$

Joint Distribution:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{12}$$

Posterior Inference (Intractable):

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{13}$$

Variational Inference for LDA

Variational Lower Bound (ELBO)
Approximate posterior: $q(\theta, z|\gamma, \phi)$

$$\log p(w|\alpha, \beta) \geq \mathcal{L}(\gamma, \phi; \alpha, \beta) = \mathbb{E}_q[\log p(\theta, z, w|\alpha, \beta)] - \mathbb{E}_q[\log q(\theta, z)] \tag{14}$$

Mean Field Approximation:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{15}$$

Update Equations:

$$\phi_{ni} \propto \beta_{iw_n} \exp(\Psi(\gamma_i) - \Psi(\sum_j \gamma_j)) \tag{16}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{17}$$

where Ψ is the digamma function.

Neural Topic Models: VAE-LDA

Variational Autoencoder for Topics
Encoder (inference network):

$$q_\phi(\theta|d) = \mathcal{N}(\mu_\phi(d), \Sigma_\phi(d)) \tag{18}$$

Decoder (generative network):

$$p_\psi(w|\theta) = \prod_{n=1}^{N} \text{Softmax}(W\theta + b)_{w_n} \tag{19}$$

ELBO Objective:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\theta|d)}[\log p_\psi(d|\theta)] - D_{KL}(q_\phi(\theta|d)||p(\theta)) \tag{20}$$

Reparameterization Trick:

$$\theta = \mu_\phi(d) + \epsilon \odot \sigma_\phi(d), \quad \epsilon \sim \mathcal{N}(0, I) \tag{21}$$

Algorithm Pipeline

1. **Document Embeddings:** $e_i = \text{BERT}(d_i)$
2. **Dimensionality Reduction:** UMAP

$$\mathcal{L}_{UMAP} = \sum_{i \sim j} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}} \tag{22}$$

3. **Clustering:** HDBSCAN with mutual reachability
4. **Topic Representation:** c-TF-IDF

$$w_{t,c} = tf_{t,c} \cdot \log \left(1 + \frac{|C|}{|\{c' : t \in c'\}|} \right) \tag{23}$$

where $tf_{t,c}$ is term frequency in cluster c.

Dynamic Topic Models

Time-Evolving Topics
State evolution: $\beta_{k,t}|\beta_{k,t-1} \sim \mathcal{N}(\beta_{k,t-1}, \sigma^2 I)$
Online LDA with Mini-Batch Updates:

$$\rho_t = (\tau_0 + t)^{-\kappa}, \quad \lambda_t = (1 - \rho_t)\lambda_{t-1} + \rho_t \tilde{\lambda}_t \tag{24}$$

Change Point Detection: Bayesian online changepoint detection with hazard function $H(r)$. Predictive probability:

$$P(x_t|x_{1:t-1}) = \sum_{r=0}^{t-1} P(r_t = r|x_{1:t-1})P(x_t|x_{r+1:t-1}) \tag{25}$$

Temporal Coherence: Regularization term $\Omega = \sum_t ||\beta_t - \beta_{t-1}||^2$

Hierarchical Dirichlet Process

$$G_j|G_0 \sim DP(\alpha, G_0), \quad G_0|\gamma, H \sim DP(\gamma, H) \tag{26}$$

Chinese Restaurant Process: Customer n sits at table k with probability:

$$P(\text{table } k) = \begin{cases} \frac{n_k}{n-1+\alpha} & \text{if } k \text{ occupied} \\ \frac{\alpha}{n-1+\alpha} & \text{if new table} \end{cases} \tag{27}$$

Pachinko Allocation Model: Topic correlations via directed acyclic graph (DAG).
Tree-Structured Topic Hierarchies: Nested partitions with depth-dependent Dirichlet parameters.

Topic Coherence and Evaluation

PMI Coherence:

$$C_{PMI} = \frac{2}{k(k-1)} \sum_{i<j} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{28}$$

Normalized PMI (NPMI):

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log P(w_i, w_j)} \tag{29}$$

CV Coherence: Sliding window with normalized pointwise mutual information.
Word Embedding Coherence:

$$C_{emb} = \frac{1}{k(k-1)} \sum_{i\neq j} \text{sim}(\text{embed}(w_i), \text{embed}(w_j)) \tag{30}$$

Evaluation: Coherence correlates with human interpretability judgments.

From Headlines to Narratives: Aggregation Theory

Graph-Based Headline Clustering

Similarity Graph Construction
Edge weights between headlines h_i, h_j:

$$w_{ij} = \text{sim}(h_i, h_j) = \frac{\text{BERT}(h_i) \cdot \text{BERT}(h_j)}{||\text{BERT}(h_i)|| \cdot ||\text{BERT}(h_j)||} \tag{31}$$

Spectral Clustering Objective:

$$\min_H \text{tr}(H^T L H) \quad \text{s.t.} \quad H^T H = I \tag{32}$$

where $L = D - W$ is the graph Laplacian.
Solution: Eigenvectors of smallest eigenvalues of L
Normalized Cut:

$$\text{NCut}(A, B) = \frac{\text{cut}(A, B)}{\text{vol}(A)} + \frac{\text{cut}(A, B)}{\text{vol}(B)} \tag{33}$$

Multi-Document Summarization

Integer Linear Programming Formulation
Binary variables: $x_i = 1$ if sentence i selected
Objective:

$$\max \sum_i w_i x_i - \lambda \sum_{i,j} s_{ij} x_i x_j \tag{34}$$

where w_i is importance, s_{ij} is similarity.
Constraints:

$$\sum_i l_i x_i \leq L \quad \text{(length limit)} \tag{35}$$

$$\sum_{i \in C_k} x_i \geq 1 \quad \forall k \quad \text{(coverage)} \tag{36}$$

Submodular Approximation:

$$f(S) = \sum_{i \in V} \min(R(i,S), \alpha R(i,V)) \tag{37}$$

Greedy algorithm gives $(1 - 1/e)$ approximation.

Event Chain Extraction

Narrative Event Chains
Probability of event sequence:

$$P(e_1, ..., e_n) = P(e_1) \prod_{i=2}^{n} P(e_i | e_1, ..., e_{i-1}) \tag{38}$$

Pairwise Event Relations:
Temporal: *before, after, simultaneous*
Causal: $P(\text{cause}(e_i, e_j) | e_i, e_j, \text{context})$
Script Learning Objective:

$$\max_\theta \sum_{(e_i, e_j) \in \mathcal{D}} \log P_\theta(e_j | e_i) + \log P_\theta(\text{rel}_{ij} | e_i, e_j) \tag{39}$$

Coherence Score:

$$\text{coherence}(e_1, ..., e_n) = \prod_{i<j} P(e_j | e_i)^{1/d_{ij}} \tag{40}$$

Cross-Document Entity Linking

Entity Resolution Across Documents
Mention pair scoring:

$$s(m_i, m_j) = w^T \phi(m_i, m_j, \text{context}) \tag{41}$$

Clustering Objective:

$$\max \sum_{C \in \mathcal{C}} \sum_{m_i, m_j \in C} s(m_i, m_j) - \lambda ||\mathcal{C}|| \tag{42}$$

Knowledge Graph Construction:
Nodes: Entities $\mathcal{E}$, Events $\mathcal{V}$
Edges: Relations $\mathcal{R}$
Narrative Graph:

$$G = (\mathcal{E} \cup \mathcal{V}, \mathcal{R}), \quad \mathcal{R} \subseteq (\mathcal{E} \times \mathcal{V}) \cup (\mathcal{V} \times \mathcal{V}) \tag{43}$$

Dempster-Shafer Theory for Evidence Combination
Basic probability assignment: $m : 2^\Theta \to [0, 1]$
Belief and Plausibility:

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B) \tag{44}$$

$$\text{Pl}(A) = \sum_{B \cap A \neq \emptyset} m(B) \tag{45}$$

Dempster's Rule of Combination:

$$m_{12}(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K} \tag{46}$$

where $K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ is the conflict.
Contradiction Detection: High conflict K indicates inconsistent sources.

Narrative Arc Modeling

Freytag's Pyramid Formalization
Dramatic arc structure: Exposition → Rising Action → Climax → Falling Action → Denouement
Sentiment Trajectory:

$$s(t) = \sum_i w_i \cdot \text{sentiment}(w_i, t) \tag{47}$$

Vonnegut's Story Shapes:

- Man in Hole: $s(t) = -\sin(\pi t) + \epsilon(t)$
- Cinderella: $s(t) = \text{sigmoid}(\alpha(t - t_0)) + \epsilon(t)$
- Kafka: $s(t) = -e^{-\lambda t} + \epsilon(t)$

Story Grammar Parsing: Context-free grammar: $S \rightarrow$ Setup Complication Resolution
Neural Implementation: RNN/Transformer with narrative structure attention.

Transformer Architecture for Narrative Understanding

Multi-Head Self-Attention Mechanism

Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{48}$$

Multi-Head Attention:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{49}$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{50}$$

Complexity Analysis:

- Self-attention: $O(n^2 \cdot d)$
- Feed-forward: $O(n \cdot d^2)$
- Memory: $O(n^2 + n \cdot d)$

where $n =$ sequence length, $d =$ model dimension

Absolute Position Encoding
Sinusoidal:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \tag{51}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \tag{52}$$

Relative Position Encoding:

$$e_{ij} = x_i W^Q (x_j W^K + a_{ij}^K)^T + x_i W^Q (a_{ij}^K)^T \tag{53}$$

Rotary Position Embedding (RoPE):

$$f_q(x_m, m) = R_{\Theta,m}^d W_q x_m \tag{54}$$

where $R_{\Theta,m}^d$ is a rotation matrix dependent on position m.

Sparse Attention Patterns

Longformer sliding window + global:

$$\text{Attention}_{ij} = \begin{cases} 1 & \text{if } |i-j| \leq w/2 \\ 1 & \text{if } i \in \mathcal{G} \text{ or } j \in \mathcal{G} \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

Linear Attention via Kernel Trick:

$$\text{Attention}(Q, K, V) = \phi(Q)(\phi(K)^T V) \tag{56}$$

Complexity: $O(n \cdot d^2)$ instead of $O(n^2 \cdot d)$

Flash Attention:

- Tiling computation for GPU memory hierarchy
- Recomputation in backward pass
- IO complexity: $O(n^2 d / M^{1/2})$

Pre-training Objectives for Narrative Understanding

Beyond MLM: Discourse-Aware Objectives
Sentence Order Prediction (SOP):

$$\mathcal{L}_{SOP} = -\mathbb{E}_{(s_1,s_2)} \log P(y|[CLS], s_1, [SEP], s_2) \tag{57}$$

Discourse Relation Prediction:

$$\mathcal{L}_{DR} = -\sum_{r\in\mathcal{R}} \log P(r|s_i, s_j, \text{context}) \tag{58}$$

Contrastive Learning (SimCSE):

$$\mathcal{L} = -\log \frac{e^{\text{sim}(h_i,h_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i,h_j)/\tau}} \tag{59}$$

where h_i^+ is augmented version of h_i, τ is temperature.

Fine-tuning Strategies for Narratives

Parameter-Efficient Fine-tuning
LoRA (Low-Rank Adaptation):

$$W' = W + BA, \quad B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k} \tag{60}$$

Adapters:

$$h' = h + f(hW_{down})W_{up} \tag{61}$$

Prefix-tuning:

$$P(y|x) = P(y|P_\phi; x) \tag{62}$$

Multi-task Learning:

$$\mathcal{L}_{total} = \sum_{i=1}^{T} w_i \mathcal{L}_i + \lambda\Omega(\theta) \tag{63}$$

Advantage: Efficient adaptation while preserving pre-trained knowledge.

Attention Analysis for Narrative Structure
Attention rollout:

$$A_{rollout} = \prod_{l=1}^{L} A^{(l)} \tag{64}$$

Probing Tasks:

- Syntactic: POS tagging, dependency parsing
- Semantic: Named entity recognition, coreference
- Discourse: Narrative structure, coherence

Gradient-based Attribution: Integrated gradients:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{65}$$

Tools: BertViz, Captum, Attention rollout for narrative flow analysis.

Large Language Models for Narrative Generation

Autoregressive Language Modeling

Fundamental Objective

$$P(x_1, ..., x_T) = \prod_{t=1}^{T} P(x_t|x_{<t}) \tag{66}$$

Maximum Likelihood Training:

$$\mathcal{L}_{MLE} = -\frac{1}{T}\sum_{t=1}^{T} \log P_\theta(x_t|x_{<t}) \tag{67}$$

Perplexity:

$$\text{PPL} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log P(x_t|x_{<t})\right) = \exp(\mathcal{L}_{MLE}) \tag{68}$$

Exposure Bias Problem:

- Training: Teacher forcing with ground truth
- Inference: Model's own predictions
- Mismatch leads to error accumulation

Scheduled Sampling: Use model predictions with probability ϵ_t:

$$x_t^{input} = \begin{cases} x_t^{truth} & \text{with prob } 1 - \epsilon_t \\ \arg\max P(x|x_{<t}) & \text{with prob } \epsilon_t \end{cases} \tag{69}$$

Beam Search
Maintain top k sequences:

$$\hat{y} = \arg\max_y P(y|x) = \arg\max_y \prod_{t=1}^{T} P(y_t|y_{<t}, x) \tag{70}$$

Top-k Sampling:

$$P'(x) = \begin{cases} P(x)/\sum_{x' \in V_k} P(x') & \text{if } x \in V_k \\ 0 & \text{otherwise} \end{cases} \tag{71}$$

Nucleus (Top-p) Sampling:

$$V_p = \text{smallest set s.t.} \sum_{x \in V_p} P(x) \geq p \tag{72}$$

Temperature Scaling:

$$P'(x_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{73}$$

Controlled Generation with LLMs

Plug and Play Language Models (PPLM)
Gradient-based steering:

$$\tilde{H}_t = H_t + \alpha \frac{\nabla_{H_t} \log P(a|x + H_t)}{\||\nabla_{H_t} \log P(a|x + H_t)||_\gamma} \tag{74}$$

Control Codes (CTRL):

$$P(x|c) = \prod_{t=1}^{T} P(x_t|c, x_{<t}) \tag{75}$$

Reinforcement Learning from Human Feedback:
Reward model: $r_\phi(x, y)$
Policy optimization:

$$\mathcal{L}_{RLHF} = -\mathbb{E}_{y\sim\pi_\theta}[r_\phi(x, y)] + \beta D_{KL}(\pi_\theta||\pi_{ref}) \tag{76}$$

In-Context Learning Theory

Few-Shot Learning via Prompting
Prompt structure: $P = (x_1, y_1, ..., x_k, y_k, x_{test})$
Bayesian Interpretation:

$$P(y|x, \mathcal{D}_{prompt}) = \int P(y|x, \theta) P(\theta|\mathcal{D}_{prompt}) d\theta \tag{77}$$

Implicit Meta-Learning:

$$\theta^* = \arg\min_\theta \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} [\mathcal{L}(\theta, \mathcal{T}_{support})] \tag{78}$$

Gradient Descent in Context: LLMs perform implicit gradient descent on in-context examples.
Distribution Shift:

$$\text{Error} \leq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + D_{TV}(P_{train}, P_{test})\right) \tag{79}$$

Hallucination Detection
Confidence-based detection:

$$H(x) = 1 - \max_y P(y|x) \tag{80}$$

Factual Consistency Scoring:

$$FC(s, d) = \frac{1}{|E_s|} \sum_{e \in E_s} \mathbb{I}[e \text{ entailed by } d] \tag{81}$$

Knowledge Grounding: Retrieval-augmented generation (RAG):

$$P(y|x) = \sum_{z \in \text{retrieve}(x)} P(z|x)P(y|x, z) \tag{82}$$

Uncertainty Quantification: Semantic entropy: $SE = -\sum_c P(c|x) \log P(c|x)$
Applications: Filter unreliable narrative generations.

Evaluation Metrics for Narrative Generation

Overlap-Based Metrics
BLEU score:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{83}$$

ROUGE-L:

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{\beta^2 R_{lcs} + P_{lcs}} \tag{84}$$

Semantic Similarity:
BERTScore:

$$F_{BERT} = \frac{2 \cdot P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{85}$$

Human Evaluation: Coherence, fluency, informativeness, factuality

Advanced NLP Theory for Narratives

End-to-End Neural Model

Span representation:

$$g_i = [x_{START(i)}, x_{END(i)}, \hat{x}_i, \phi(i)] \tag{86}$$

Mention scoring:

$$s_m(i) = w_m^T \text{FFNN}_m(g_i) \tag{87}$$

Pairwise scoring:

$$s_a(i,j) = w_a^T \text{FFNN}_a([g_i, g_j, g_i \circ g_j, \phi(i,j)]) \tag{88}$$

Marginalization over antecedents:

$$P(y_i = j) = \frac{\exp(s(i,j))}{\sum_{j' \in Y(i)} \exp(s(i,j'))} \tag{89}$$

where $s(i,j) = s_m(i) + s_m(j) + s_a(i,j)$

Allen's Interval Algebra
13 basic relations: before, after, meets, overlaps, starts, finishes, equals...
Temporal Graph Construction:
Nodes: Events $\mathcal{E}$
Edges: Temporal relations $\mathcal{R}$
Constraint Propagation:

$$r_{AC} = r_{AB} \circ r_{BC} \tag{90}$$

TimeML Annotation:

- EVENT: Actions, states
- TIMEX3: Temporal expressions
- SIGNAL: Temporal connectives
- TLINK: Temporal links

Neural Temporal Extraction:

$$P(r_{ij}|e_i, e_j) = \text{softmax}(W[\text{BERT}(e_i); \text{BERT}(e_j); \text{features}]) \tag{91}$$

Relation Extraction for Narratives

Distant Supervision
Automatically label training data using knowledge base:

$$\mathcal{L}_{distant} = \sum_{(e_1, r, e_2) \in KB} \log P(r | \text{sentences containing } e_1, e_2) \tag{92}$$

Joint Entity and Relation Extraction:

$$P(E, R|S) = P(E|S) \cdot P(R|E, S) \tag{93}$$

Graph Neural Networks for RE:

$$h_i^{(l+1)} = \sigma\left(\sum_{r \in \mathcal{R}} \sum_{j \in N_r(i)} W_r^{(l)} h_j^{(l)} + b^{(l)}\right) \tag{94}$$

OpenIE: Open Information Extraction Extract $(arg_1, relation, arg_2)$ tuples without predefined schema.

Causal Reasoning in Text

Pearl's Causal Hierarchy in NLP
Level 1 - Association: $P(Y|X)$ Statistical correlation in text.
Level 2 - Intervention: $P(Y|do(X))$ Counterfactual text generation.
Level 3 - Counterfactuals: $P(Y_x|X', Y')$ What would have happened if...
Backdoor Adjustment:

$$P(Y|do(X)) = \sum_z P(Y|X, Z)P(Z) \tag{95}$$

Instrumental Variables in Text: Use exogenous text features as instruments for causal identification.
Application: Distinguish causation from correlation in narrative claims.

Aspect-Based Sentiment Analysis

Joint Extraction and Classification
Task: Extract aspects and predict sentiment
BERT for ABSA:

$$h = \text{BERT}([CLS], \text{sentence}, [SEP], \text{aspect}, [SEP]) \tag{96}$$

CRF Layer for Sequence Labeling:

$$P(y|x) = \frac{\exp(\sum_{i=1}^{n}(W_{y_{i-1},y_i} + P_{i,y_i}))}{\sum_{y'}\exp(\sum_{i=1}^{n}(W_{y'_{i-1},y'_i} + P_{i,y'_i}))} \tag{97}$$

Multi-Task Learning:

$$\mathcal{L} = \lambda_1\mathcal{L}_{aspect} + \lambda_2\mathcal{L}_{sentiment} + \lambda_3\mathcal{L}_{joint} \tag{98}$$

Mathematical Optimization for NLP

Beyond Cross-Entropy
Focal Loss (for imbalanced data):

$$\mathcal{L}_{focal} = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{99}$$

Label Smoothing:

$$y'_i = (1 - \epsilon) y_i + \frac{\epsilon}{K} \tag{100}$$

Contrastive Loss (InfoNCE):

$$\mathcal{L}_{NCE} = -\log \frac{\exp(f(x, x^+)/\tau)}{\exp(f(x, x^+)/\tau) + \sum_{i=1}^{N} \exp(f(x, x_i^-)/\tau)} \tag{101}$$

Sequence-Level Loss:

$$\mathcal{L}_{seq} = -\mathbb{E}_{y \sim P_\theta}[R(y)] + \lambda H(P_\theta) \tag{102}$$

Optimization Algorithms for Transformers

Adam with Warmup
Adam update:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{103}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{104}$$

$$\theta_t = \theta_{t-1} - \eta \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{105}$$

Learning Rate Schedule:

$$\eta_t = d_{model}^{-0.5} \cdot \min(t^{-0.5}, t \cdot \text{warmup}^{-1.5}) \tag{106}$$

Gradient Clipping:

$$g' = \begin{cases} g & \text{if } ||g|| \leq c \\ c \cdot g/||g|| & \text{otherwise} \end{cases} \tag{107}$$

Regularization Techniques

Dropout

$$y = \frac{1}{1-p} x \odot m \text{ where } m \sim \text{Bernoulli}(1-p) \tag{108}$$

Layer Normalization:

$$y = \gamma \frac{x-\mu}{\sigma} + \beta \tag{109}$$

Weight Decay (L2 Regularization):

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \sum_i ||W_i||_2^2 \tag{110}$$

Spectral Normalization:

$$W_{SN} = \frac{W}{\sigma(W)}, \quad \sigma(W) = \max_u ||Wu||_2 \tag{111}$$

where $\sigma(W)$ is the spectral norm (largest singular value).

Multi-objective Optimization

Pareto Optimality in Multi-task Learning
Pareto frontier: No improvement in one task without degrading another.
Gradient Surgery:

$$g_i' = g_i - \frac{g_i \cdot g_j}{||g_j||^2} g_j \text{ if } g_i \cdot g_j < 0 \tag{112}$$

Uncertainty Weighting:

$$\mathcal{L} = \sum_i \frac{1}{2\sigma_i^2} \mathcal{L}_i + \log \sigma_i \tag{113}$$

Task Balancing: Dynamic loss scaling: $w_i^{(t)} = \frac{r_i^{(t-1)}}{\sum_j r_j^{(t-1)}}$

where $r_i^{(t)}$ is the loss ratio for task i.

Information Theory for Narrative Analysis

Mutual Information between Headlines and Narratives

$$I(H;N) = \sum_{h,n} P(h,n) \log \frac{P(h,n)}{P(h)P(n)} = H(N) - H(N|H) \tag{114}$$

Conditional Entropy:

$$H(N|H) = -\sum_{h,n} P(h,n) \log P(n|h) \tag{115}$$

Information Gain:

$$IG = H(N) - \sum_{h} P(h) H(N|H=h) \tag{116}$$

Jensen-Shannon Divergence:

$$JS(P,Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M), \quad M = \frac{P+Q}{2} \tag{117}$$

Application: Quantify information content in narrative extraction.

Spectral Theory for Attention Analysis

Eigendecomposition of Attention Matrices
Attention matrix: $A = QK^T/\sqrt{d_k}$
Spectral Decomposition:

$$A = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T \tag{118}$$

Low-rank Approximation:

$$A_k = \sum_{i=1}^{k} \lambda_i u_i u_i^T \tag{119}$$

Attention Head Analysis: Rank of attention head: $\text{rank}(A_h) = |\{\lambda_i : \lambda_i > \epsilon\}|$
Frobenius Norm:

$$||A||_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\text{tr}(A^T A)} = \sqrt{\sum_i \lambda_i^2} \tag{120}$$

Insight: Low-rank attention heads focus on specific linguistic patterns.

Key Mathematical Contributions

Unified Framework for Narrative Extraction

1. **Embedding Theory:** Distributional semantics $\rightarrow$ contextual representations
2. **Topic Discovery:** Probabilistic graphical models $\rightarrow$ neural variational inference
3. **Aggregation:** Graph algorithms + optimization for multi-document fusion
4. **Transformer Architecture:** Attention as information routing
5. **Generation Theory:** Autoregressive models with control
6. **Advanced NLP:** Joint models for narrative understanding
7. **Optimization:** Specialized techniques for language models

Open Questions:

- Causal representation learning in text
- Compositional generalization
- Grounded language understanding

Thank You

Questions and Discussion

Contact:
Prof. Dr. Joerg Osterrieder

Resources:
Code implementations available

Mathematical proofs in supplementary materials