Topic-Aware Text Summarization

Presenter: Yu Yang

School of Statistics
University of Minnesota

April 21, 2021

Why Text Summarization?

- Reduce information overload.
- Get the main points without reading the entire document
- Save time and effort
- Widely used: News, books, biomedical documents, legal documents, and scientific papers.

Categorization

- Based on summarization approach
  - extractive
  - abstractive
- Based on summary type
  - Headline
  - Sentence-level
  - Highlights
  - Full Summary

Text Summarization vs. Machine Translation

- Both are sequence-to-sequence problems.
- Machine translation has a **one-to-one** semantic correspondence between source and target words.
- Text summarization does the mapping in a **lossy** manner.

Why Incorporating Topic Representation?

- Intuition: human tend to summarize under sub-categories and then combine them together to form the final summary.
- Information: reveals more global semantic information and captures corpus-level patterns of words co-occurrence.
- Higher level of abstraction: might capture the lossy mapping pattern.
- Usage: might provide valuable inductive bias for language generation models.

Topic Modeling Methods

- Latent Dirichlet Allocation (LDA)
- Replicated Softmax (a family of Restricted Boltzmann Machines)
- Deep Boltzmann Machine (DBM)
- Variational Autoencoder (VAE)

Goals of this Project

- Not target at beating the state-of-the-art models currently.
- Investigate appropriate ways to incorporate topic representations.
- Compare with the baseline models and check the effect of inserting topic information.

Basic Model Framework

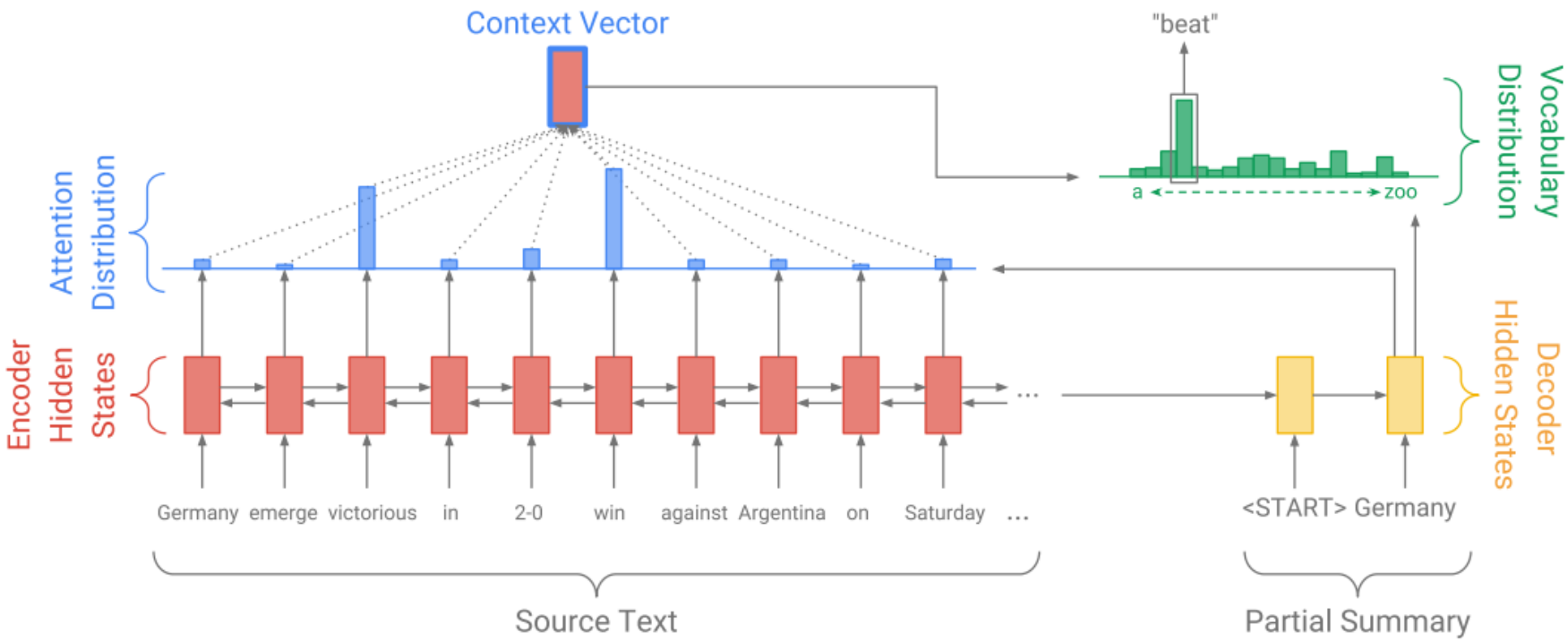


Figure 1: Baseline sequence-to-sequence model with attention

Pointer-generator Model (See et al., 2017)

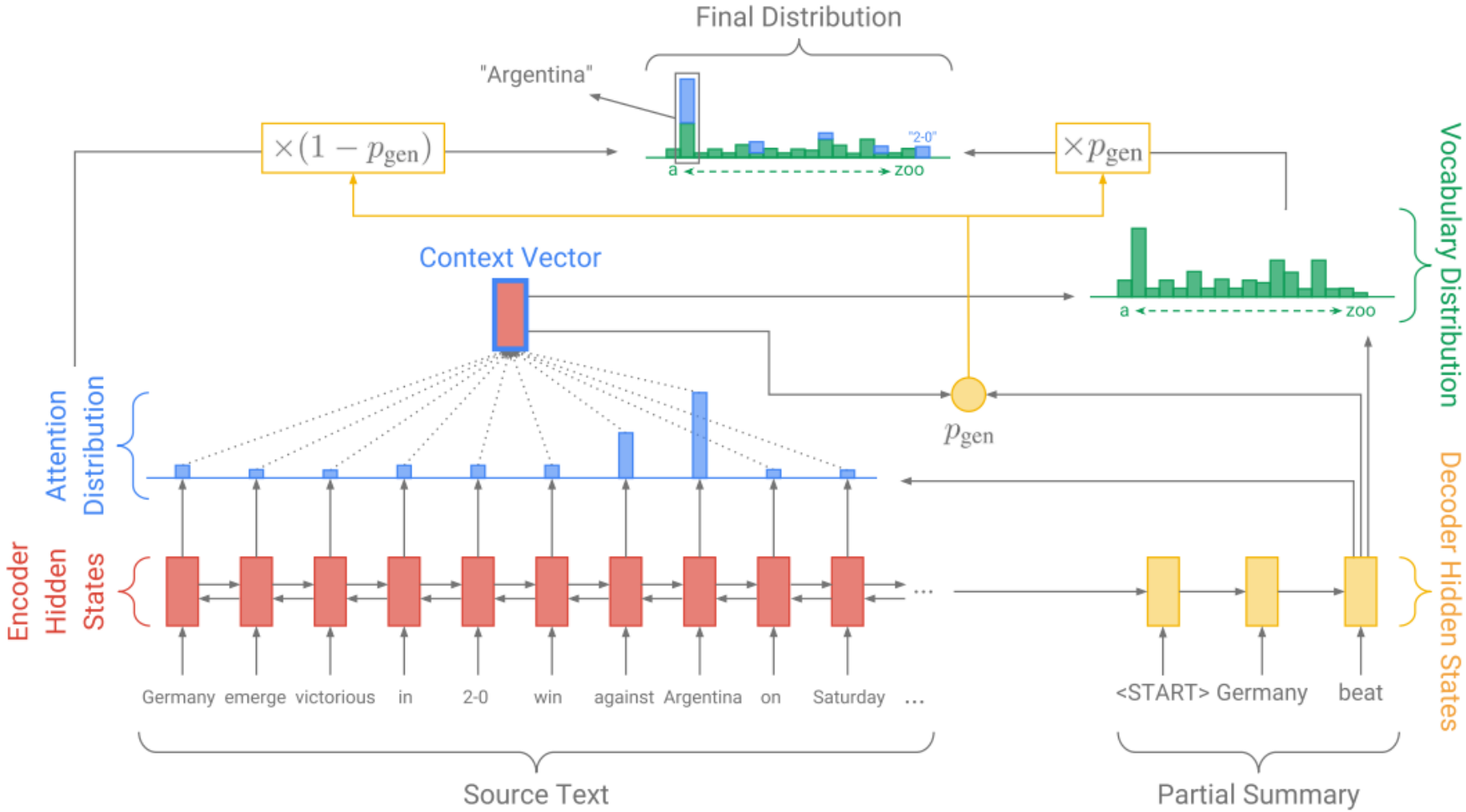


Figure 2: Pointer-generator Model (To enable copy mechanism.)

Replicated Softmax (Hinton and Salakhutdinov, 2009)

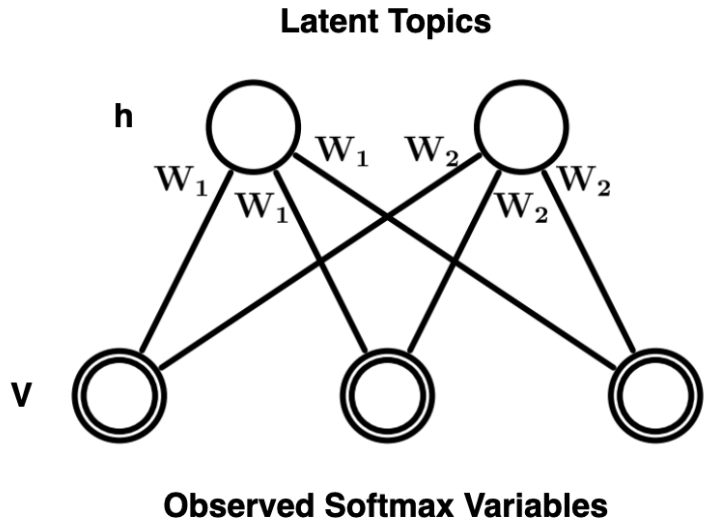


Figure 3: Replicated Softmax

Energy function:

$$E(\mathrm{V},\mathrm{h}) = -\sum_{j=1}^{F}\sum_{k=1}^{K} W_j^k h_j \hat{v}^k - \sum_{k=1}^{K} \hat{v}^k b^k - D\sum_{j=1}^{F} h_j a_j,$$

where $\hat{v}^k = \sum_{i=1}^{D} v_i^k$ denotes the count for the kth word.

Proposal: Topic-Aware Text Summarizer

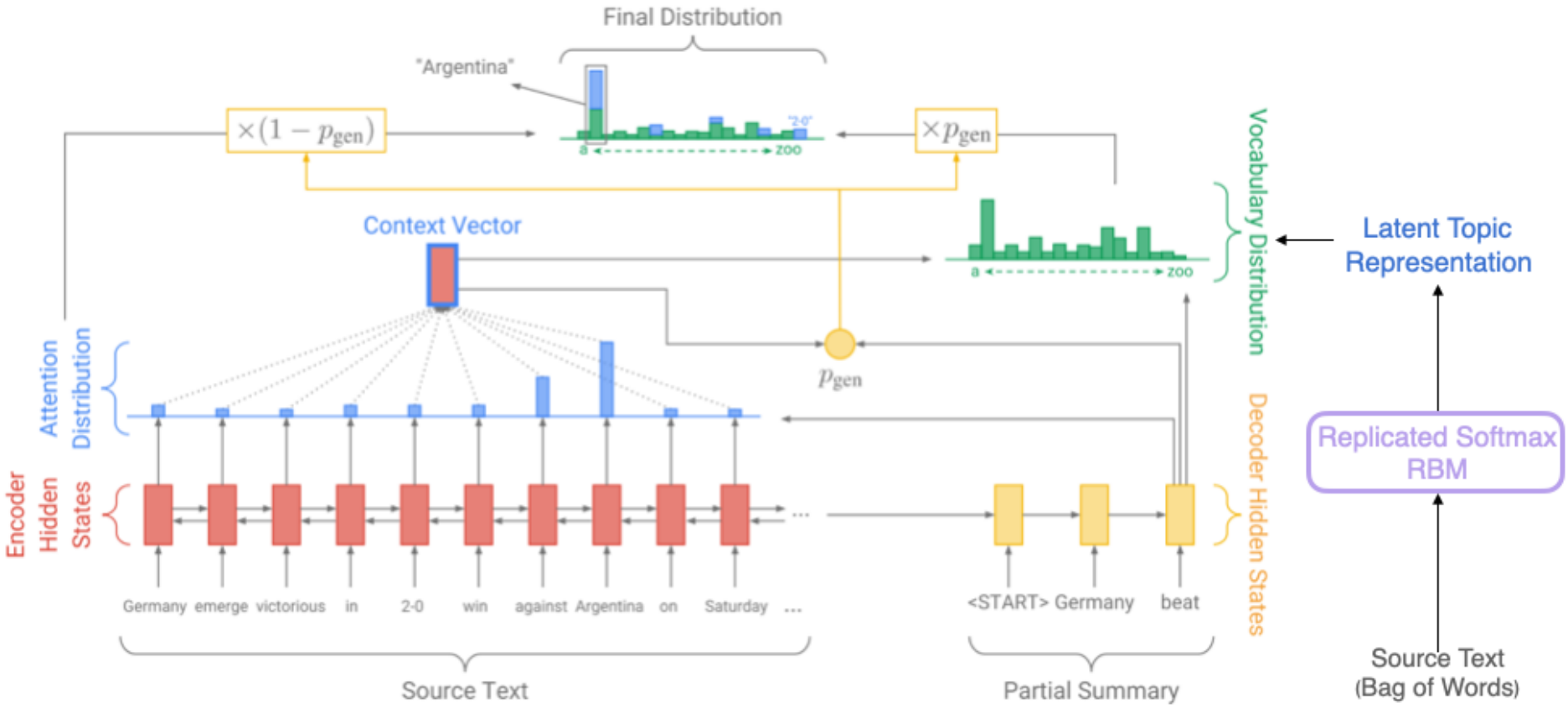


Figure 4: Topic-Aware Text Summarizer: PGNet + Text RBM

Formulation

- Notation: encoder hidden state h_i, decoder input x_t, decoder hidden state s_t, target word w_t^*, and latent topic representation l.
- attention: $e_i^t = v^T \tanh\left(W_h h_i + W_s s_t + b_{\text{attn}}\right)$, $a^t = \text{softmax}\left(e^t\right)$
- context vector: $h_t^* = \sum_i a_i^t h_i$
- generation probability: $p_{\text{gen}} = \sigma\left(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}}\right)$
- vocabulary distribution

$$P_{\text{vocab}} = \text{softmax}\left(V'\left(V\left[s_t, h_t^*\right] + b\right) + b' + WI\right)$$

- final distribution over the extended vocabulary

$$P(w) = p_{\text{gen}}\ P_{\text{vocab}}\ (w) + (1 - p_{\text{gen}}\ ) \sum_{i:w_i=w} a_i^t$$

- loss $= \frac{1}{T}\sum_{t=0}^{T} \text{loss}_t = -\frac{1}{T}\sum_{t=0}^{T} \log P\left(w_t^*\right)$

Dataset

CNN/Daily Mail dataset (Nallapati et al., 2016)

- contains online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average).
- We used the same version of the data as See et al., 2017: non-anonymized, processed by Stanford CoreNLP tool, and has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs.

Original Text: (CNN) – An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesday, the same ship on which 86 passengers previously fell ill, according to the state-run Brazilian news agency, Agencia Brasil. The American tourist died aboard the MS Veendam, owned by cruise operator Holland America. Federal Police told Agencia Brasil that forensic doctors were investigating her death. The ship's doctors told police that the woman was elderly and suffered from diabetes and hypertension, according the agency. The other passengers came down with diarrhea prior to her death during an earlier part of the trip, the ship's doctors said. The Veendam left New York 36 days ago for a South America tour.
Ground Truth Summary: The elderly woman suffered from diabetes and hypertension, ship's doctors say .Previously, 86 passengers had fallen ill on the ship, Agencia Brasil says .

Figure 5: CNN/DailyMail Dataset Example

Training

Replicated Softmax:

- Preprocessing: remove common stopwords, do lemmatization, and keep 15,000 most frequent words in the training dataset.
- Latent dimension: 200
- Contrastive Divergence training: k = 1 (Ideally, should increase k from 1 to 5 gradually through the training process).

PGNet + RBM:

- Pretrain the RBM model to get latent representation vectors for the documents.
- Hidden dimension: 256, embedding dimension: 128, vocabulary size: 50,000, beam size: 4.
- Train the modified PGNet with batch size 8, iteration steps 500,000, and optimizer Autograd with learning rate 0.15 and an initial accumulator value of 0.1.

Results

Table 1: ROUGE F_1 scores on the test set

Method	ROUGE[1]		
	1	2	L
PGNet (original paper)	36.44	15.66	33.42
PGNet (my 500k run)	35.92 (35.70, 36.16)	15.51 (15.31, 15.73)	32.87 (32.65, 33.11)
PGNet (my 100k run)	**36.98** (36.74, 37.21)	**15.93** (15.72, 16.15)	**33.55** (33.32, 33.77)
PGNet + Text RBM (my 100k run)	36.33 (36.10, 36.56)	15.62 (15.41, 15.84)	33.05 (32.82, 33.28)

Training Curve on Tensorboard

[1]ROUGE-N: Overlap of N-grams between the system and reference summaries. ROUGE-L: Longest Common Subsequence based statistics.

Future Work & Lessons

Future Work

1. Examine the quality of the learned topic representation.
2. Use attention mechanism to insert the topic vectors.
3. Do qualitative analysis to examine where the proposed model goes wrong.

Some Lessons about Coding

1. Check the datasets first: availability, processing, and loading.
2. Pay attention to bias terms to avoid over-parameterization.
3. Pay attention to overflow and underflow issues when there are exponential operation.
4. Run the model for a few iterations and check the output.

References

Hinton, Geoffrey E and Russ R Salakhutdinov (2009). "Replicated softmax: an undirected topic model". In: *Advances in neural information processing systems* 22, pp. 1607–1614.

Nallapati, Ramesh et al. (2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290.

See, Abigail, Peter J Liu, and Christopher D Manning (2017). "Get To The Point: Summarization with Pointer-Generator Networks". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083.

Improvement Tracks

Prior to 2018

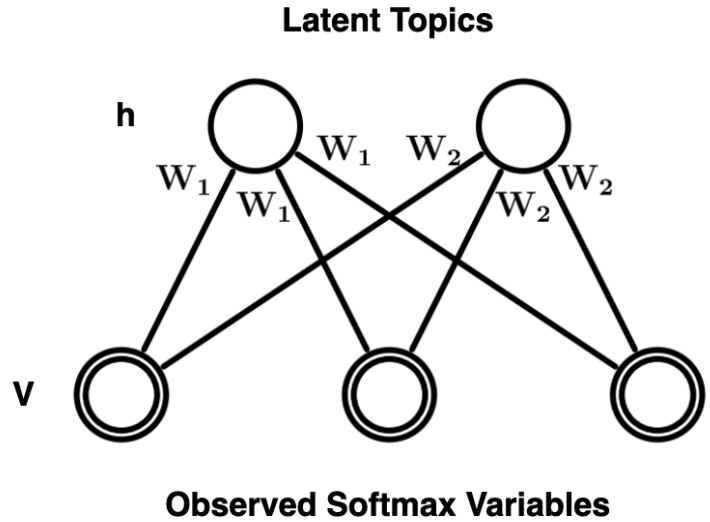
- Input: NLP feature vectors, contextual embeddings, topic distribution vectors...
- Encoder/Decoder: RNN, LSTM, CNN, Transformer, ...
- Attention: hierarchical, graph-based, ...
- Decoding Mechanism: copy mechanism, ...

After 2018

Pre-training Transformer-based models: self-supervised objectives, masking strategies, parameter sharing, model size, training tricks, ...

- Pre-training specifically for text summarization
- Pre-training for sequence-to-sequence models in general

Replicated Softmax II



Conditional distributions:

$$p\left(v_i^k = 1 \mid \mathrm{h}\right) = \frac{\exp\left(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k\right)}{\sum_{q=1}^{K} \exp\left(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q\right)}$$

$$p\left(h_j = 1 \mid \mathrm{V}\right) = \sigma\left(Da_j + \sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k W_{ij}^k\right)$$