Morphology-Aware Multi-Granularity Representation Learning for Agglutinative Languages

Zhonghao Zhang[1], Na Liu[1,*], Jiajia Ma[1], Nier WU[1], Guiping Liu[1]
[1]Inner Mongolia University of Technology
{20241800131,csnaliu,20241800142,wunier04,csguipingliu}@imut.edu.cn
Correspondence: csnaliu@imut.edu.cn

Abstract

Low-resource agglutinative languages, characterized by rich morphological inflection and severe vocabulary sparsity in corpora, have long posed numerous challenges in the field of representation learning.Word-level representations preserve semantic integrity but struggle to handle sparse surface forms, whereas morpheme-level representations, though easier to learn, often lack holistic semantic information. Existing multi-granularity methods are typically modeled at the word and phrase levels, with very limited application to low-resource agglutinative languages. Focusing on the morphemes of agglutinative languages, this paper proposes MAGNet, a morphology-aware gated multi-granularity pre-training framework. At the morpheme granularity, this framework leverages morphological knowledge and integrates morpheme segmentation with morphological tagging to construct fine-grained representations. It further introduces a morphology-aware masked language modeling objective to facilitate the model in learning functional morphological regularities. Meanwhile, at the word granularity, a word-level encoder is employed to capture contextual semantics and maintain its semantic coherence. Finally, a gated fusion mechanism dynamically fuses representations of different granularities according to the context. Experiments conducted on two low-resource agglutinative languages, Mongolian and Turkish, for the tasks of dependency parsing and named entity recognition (NER) demonstrate that our method achieves consistent performance improvements over strong baseline models. Ablation studies further validate the complementary roles of morphological tagging and whole-word modeling in efficient representation learning.

1 Introduction

Pre-trained language models such as BERT, RoBERTa, and ALBERT (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019) have achieved state-of-the-art performance across a wide range of natural language processing tasks by leveraging large-scale pre-training on the Transformer architecture followed by task-specific fine-tuning. These models effectively capture lexical, syntactic, and semantic information in high-resource settings.

However, modeling agglutinative languages remains challenging. In such languages, words are formed by concatenating multiple morphemes that jointly encode grammatical and semantic information, resulting in a large number of surface forms and a high out-of-vocabulary rate. Especially under low-resource conditions (Liu et al., 2021), Data sparsity is further exacerbated by the combinatorial nature of morpheme sequences, limiting the effectiveness of word-level modeling.

Although contextualized models based on Transformers dynamically generate representations conditioned on context, single-granularity modeling remains insufficient for low-resource agglutinative languages: fine-grained units are easier to learn but lack holistic semantics (Arnett and Bergen, 2025), whereas coarse-grained word units preserve semantic integrity but suffer from sparsity and segmentation ambiguity (Peters et al., 2018; Devlin et al., 2019).

The development of existing multi-granularity pre-training methods has opened up an important innovative direction for natural language processing tasks. These methods have achieved remarkable performance on various benchmark tasks of high-resource languages such as English and Chinese(Sun et al., 2020; Joshi et al., 2020; Zhang et al., 2021). This technical paradigm offers direct enlightenment for tackling the modeling challenges of low-resource agglutinative languages. Nevertheless, how to design adaptive multi-granularity unit segmentation strategies, fusion mechanisms and representation learning frameworks in light of the grammatical properties of agglutinative languages

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1065–1073

constitutes the core entry point and innovative direction of this study.

In this work, we propose MAGNet (Morphology-Aware Gated multi-Granularity Pre-training), a multi-granularity pre-training framework for agglutinative languages. Our main contributions are summarized as follows:

Propose a morphology-aware multi-granularity word representation method. At the morpheme level, we construct fine-grained representations based on morphemes and their morphological annotations, and introduce a morphology-aware masked language modeling objective to help the model learn morphological rules. At the word level, we capture contextual semantic information via a word-level encoder, which effectively preserves semantic coherence and achieves multi-granularity representation.

Design a fusion method based on the gated mechanism. We introduce a learnable gated mechanism to dynamically trade off between fine-grained morphological representations and coarse-grained word-level representations, enabling the model to adaptively select the more appropriate representation granularity according to the context.

Validate the effectiveness of the proposed method in low-resource agglutinative language scenarios. Experimental results on multiple downstream tasks for Mongolian and Turkish show that the proposed multi-granularity pre-training framework outperforms various strong baseline models, verifying its effectiveness in improving the robustness and semantic expressiveness of word representations for agglutinative languages.

2 Related Work

Early studies on low-resource agglutinative languages adopted rule-based or statistical-based morphemic analysis methods, which rely heavily on extensive linguistic expertise and suffer from limited scalability (Creutz and Lagus, 2007). Subsequently, static word embedding methods such as Word2Vec and GloVe (Mikolov et al., 2013; Pennington et al., 2014) alleviated the burden of manual feature engineering, but their fixed word representation form makes it difficult to handle morphological inflection issues. Words in low-resource agglutinative languages are composed of stems concatenated with multiple affixes (Rasooli and Tetreault, 2015). To address this characteristic, the FastText(Bojanowski et al., 2017) method mitigates the representation challenges caused by morphological inflection to a certain extent by incorporating character n-gram information. However, this method still treats all subword units equally and fails to accurately capture the semantic differences and grammatical functions between stems and affixes (Heinzerling and Strube, 2018).

With the emergence of contextual representations, models such as ELMo (Peters et al., 2018), BERT, and their multilingual variants (mBERT and XLM-R) have significantly improved the effectiveness of word representation through large-scale pre-training and contextual information modeling (Devlin et al., 2019; Ruder et al., 2019). Nevertheless, in low-resource agglutinative language scenarios, these models often face problems such as poor tokenization performance and insufficient learning of rare morphological patterns, which limits their ability to fully capture fine-grained morphological semantics.

To tackle the issue of morphological richness in agglutinative languages, a series of studies have focused on explicitly integrating morphological information into word representation learning (Na et al., 2024a). These studies explored compositional models that construct representations of morphologically complex words by fusing stem and affix embeddings, demonstrating that explicit modeling of morphological structures can effectively alleviate the data sparsity problem (Lazaridou et al., 2013). Subsequent works further introduced morphological tagging or multi-task learning objectives to enable embedding vectors to capture grammatical functions while encoding semantic information (Cotterell and Schütze, 2015).

Research on multi-granularity representation learning focuses on the word and phrase levels, with the core goal of fusing information from different levels to enhance semantic representation capabilities. Among existing methods, many achieve multi-granularity fusion through simple concatenation or linear projection (Li et al., 2018). Although these methods are easy to implement, they share a common limitation: they tend to assume that all granularities contribute equally in different contextual environments. However, in low-resource agglutinative languages, the importance of morphological and lexical information carried by stems and affixes (i.e., morphemes) dynamically changes with syntactic and semantic conditions (Yang and Nicolai, 2025). Therefore, it is necessary to design

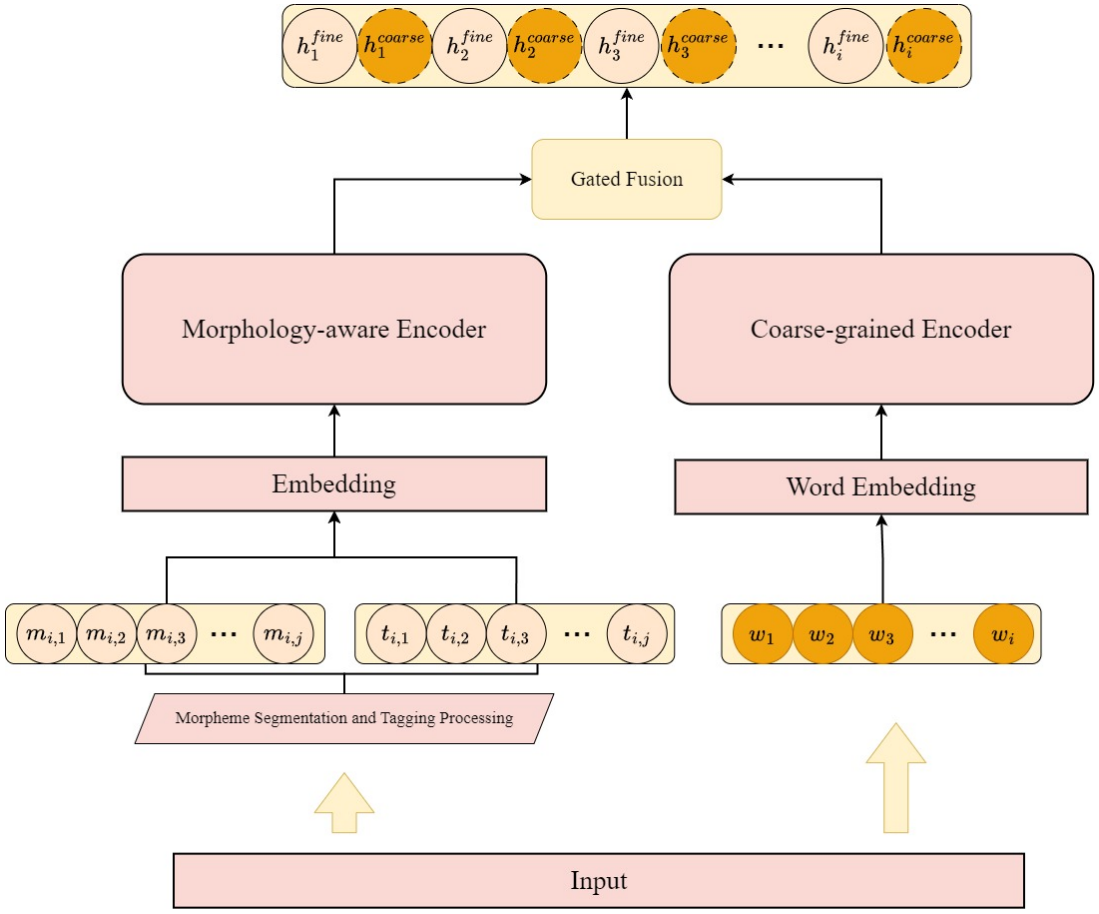

Figure 1: Overall architecture morphology-aware multi-granularity framework. The model is composed of a morphology-aware fine-grained encoder and a coarse-grained word-level encoder. After being processed in parallel via morpheme segmentation, morphological tagging, and word embeddings, the two types of representations are aligned at the word level and integrated via a gated fusion mechanism to generate the final word representations.

an adaptive fusion mechanism to dynamically balance the information weights of morphemes and whole words based on context.

3 Methodology

To simultaneously model the internal morphological structure of words and cross-word contextual semantics, we propose a unified multi-granularity encoding framework. The overall framework is shown in Figure 1, this framework constructs input sequences at different granularities and conducts contextual modeling via a shared Transformer encoder, thereby ensuring the consistency and alignability of the representation space.

3.1 Morphology-Aware Fine-Grained Representation

3.1.1 Morpheme Segmentation and Morphological Tagging

Agglutinative languages encode rich grammatical information within words through the concatenation of multiple morphemes. To explicitly model this morphological structure, this study adopts tailored schemes for Traditional Mongolian and Turkish respectively to implement morpheme segmentation and morphological tagging: for each word w_i, we first identify morpheme boundaries and segment it into an ordered morpheme sequence

$$w_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,j} \rangle \quad (1)$$

and then assign a morphological tag $t_{i,j}$ to each morpheme $m_{i,j}$ to clarify its functional role (e.g., root, case suffix, tense marker, etc.), providing structured linguistic supervision for the subsequent morphology-aware fine-grained representation learning.

For Traditional Mongolian, a pre-trained fine-tuned Multi-Task Learning Model (MTLM)(Liu et al., 2025) is directly adopted to realize the joint modeling of morpheme segmentation and morphological tagging. This method can effectively address the data sparsity problem in low-resource scenarios and avoid the error propagation problem caused by serial processing.

For Turkish, the Zemberek-NLP(Acar et al.) toolkit is employed to complete the whole process of morpheme segmentation and morphological tagging: relying on the hybrid approach of rule-based and statistical methods integrated in this toolkit, we accurately identify Turkish morpheme boundaries and perform segmentation. Meanwhile, based on the UD_Turkish-IMST(Sulubacak and Eryiğit, 2018) annotation scheme, morphological functional tags such as case, number, person, tense/mood, and voice are assigned to each segmented morpheme, directly outputting the structured morpheme sequence and corresponding tag results.

3.1.2 Fine-grained Encoding

Joint Morpheme–Tag Embedding For each morpheme $\mathbf{m}_{i,j}$, we construct a morphology-aware embedding by combining morpheme identity and morphological function:

$$\mathbf{e}_{i,j}^{\text{morph}} = \mathbf{E}_m(\mathbf{m}_{i,j}) + \mathbf{E}_t(\mathbf{t}_{i,j}) \tag{2}$$

where $\mathbf{E}_m$ and $\mathbf{E}_t$ denote morpheme and tag embedding matrices, respectively.

Morpheme-Level Encoding The sequence of joint embeddings within each word is encoded using a morphology-aware encoder:

$$\mathbf{H}_i^{\text{fine}} = \text{Encoder}_{\text{morph}}(\mathbf{e}_{i,1}^{\text{morph}}, \ldots, \mathbf{e}_{i,K_i}^{\text{morph}}) \tag{3}$$

where $\mathbf{H}_i^{\text{fine}} = \{\mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,K_i}\}$ are contextualized morpheme representations.

Intra-Word Attention Pooling To obtain a word-level fine-grained representation, we apply attention-based pooling over morphemes. The attention weight for each morpheme is computed as:

$$\alpha_{i,j} = \frac{\exp\left(\mathbf{u}^\top \tanh\left(\mathbf{W}\mathbf{h}_{i,j}\right)\right)}{\sum_{k=1}^{K_i} \exp\left(\mathbf{u}^\top \tanh\left(\mathbf{W}\mathbf{h}_{i,k}\right)\right)} \tag{4}$$

The resulting fine-grained word representation is:

$$\mathbf{h}_i^{\text{fine}} = \sum_{j=1}^{K_i} \alpha_{i,j}\, \mathbf{h}_{i,j} \tag{5}$$

This mechanism enables the model to dynamically emphasize morphemes that contribute more significantly to the semantic or grammatical meaning of the word. Such contributions are learned through the interaction between morphological supervision and contextual attention, rather than being manually assigned.

3.2 Coarse-Grained Word-Level Representation

Although the fine-grained path can explicitly model the internal morphological structure of words, it breaks the holistic semantic unit of words to a certain extent, thus affecting the modeling of semantic integrity. To this end, we introduce a parallel word-level contextual encoder to capture the semantic information of words and cross-word dependencies at the holistic level.For an input sentence:

$$S = \{w_1, w_2, \ldots, w_N\} \tag{6}$$

We first construct the word-level embedding sequence:

$$\mathbf{E}^{\text{word}} = \{\mathbf{e}_1^{\text{word}}, \mathbf{e}_2^{\text{word}}, \ldots, \mathbf{e}_N^{\text{word}}\} \tag{7}$$

$$\mathbf{e}_i^{\text{word}} = \mathbf{E}_w(w_i) + \mathbf{P}_i \tag{8}$$

where $\mathbf{E}_w(w_i)$ denotes the word embedding and $\mathbf{P}_i$ denotes the positional encoding. Different from the fine-grained path, this input does not rely on morpheme segmentation results, thus avoiding the impact of segmentation errors on semantic modeling.

3.2.1 Contextual Encoder

We employ a Transformer encoder to model the word-level sequence:

$$\mathbf{H}^{\text{coarse}} = \text{Transformer}_{\text{word}}(\mathbf{E}^{\text{word}}) \tag{9}$$

The self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{10}$$

This mechanism can capture dependencies between arbitrary pairs of words, thereby effectively modeling syntactic and semantic information.

3.2.2 Parameter Sharing and Representation Alignment

To establish representation alignment across different granularities, we adopt the following parameter strategy: the Transformer encoder parameters are shared between the fine-grained and coarse-grained paths, while the embedding layers are independent. Formally, it is expressed as:

$$\text{Transformer}_{\text{word}} \equiv \text{Transformer}_{\text{morph}} \tag{11}$$

For the i-th word, its coarse-grained representation is:

$$\mathbf{h}_i^{\text{coarse}} = \mathbf{H}_i^{\text{coarse}} \tag{12}$$

Since the fine-grained path is finally aggregated into a word-level representation $\mathbf{h}_i^{\text{fine}}$, the two paths are naturally aligned at the word granularity and can be directly used for subsequent fusion.

3.3 Gated Fusion of Multi-Granularity Representations

We propose a context-aware gated fusion mechanism to adaptively integrate information of the two granularities in a unified representation space. Benefiting from the aforementioned parameter sharing strategy, $\mathbf{h}_i^{\text{fine}}$ and $\mathbf{h}_i^{\text{coarse}}$ are already aligned in the same semantic space, enabling dimension-wise fusion.

Specifically, for the i-th word, its gating vector is defined as:

$$\mathbf{g}_i = \sigma\left(W_g\left[\mathbf{h}_i^{\text{fine}}; \mathbf{h}_i^{\text{coarse}}\right] + b_g\right) \quad (13)$$

where σ denotes the sigmoid function, $[\cdot;\cdot]$ represents the vector concatenation operation, and W_g and b_g are learnable parameters. The gating vector $\mathbf{g}_i \in (0,1)^d$ dynamically controls the contribution ratio of representations at different granularities in each dimension.

The final word representation is calculated as follows:

$$\mathbf{h}_i^{\text{final}} = \mathbf{g}_i \odot \mathbf{h}_i^{\text{fine}} + (1 - \mathbf{g}_i) \odot \mathbf{h}_i^{\text{coarse}} \quad (14)$$

where $\odot$ denotes the element-wise multiplication operation.

This mechanism allows the model to dynamically adjust the weights of the two information sources based on context: when the semantics of a word rely on its internal morphological structure (e.g., complex affixes or functional markers), the model can enhance the contribution of the fine-grained representation; when the holistic semantics are more critical (e.g., named entities or fixed expressions), it tends to rely on the coarse-grained representation.

4 Experiment

4.1 Dataset and Preprocessing

This study takes Mongolian and Turkish, two typical low-resource agglutinative languages, as its research objects. The scales of the pre-training corpora and annotated corpora for downstream tasks adopted in the experiments are as follows:

Mongolian Datasets For Mongolian, the pre-training corpus consists of 1.2 million unannotated general texts covering multiple genres such as news and folklore(Zhang et al., 2024); the dependency parsing task uses an annotated set of 20,000 sentences; for named entity recognition (NER)(S. and et al., 2016), 30,000 sentences of annotated general-domain corpora are employed, with three core entity types annotated: person names, location names, and organization names.

Turkish Datasets For Turkish, the pre-training corpus comprises 1.1 million unannotated general texts covering scenarios including news and daily communication; the dependency parsing task adopts the UD-Turkish-IMST(Sulubacak and Eryiğit, 2018) benchmark treebank, which contains 5,635 annotated sentences and over 56,000 tokens; the NER(Altinok, 2023) task utilizes the publicly available multi-genre Altinok2023 corpus.

4.2 Experimental Setup and Comparison Methods

Targeting the agglutinative features of the Mongolian language, Na et al. successively proposed the IAMC-BERT(Na et al., 2024a) and ALKI-monBERT(Na et al., 2024b) pre-trained models. By infusing agglutinative morphological knowledge into the pre-training process, these models have effectively improved the modeling performance in low-resource scenarios. However, the aforementioned studies were all conducted based on Cyrillic Mongolian, whereas the present research focuses on Traditional Mongolian . For this reason, no direct performance comparison was carried out between our model and the above-mentioned ones. Despite the discrepancies in the writing forms of the research objects, their experimental results have fully verified the effectiveness of morphological knowledge infusion in Mongolian pre-training modeling, which provides an important foundation for the method design of this study.

Although multilingual models such as mBERT (Pires et al., 2019) provide broad coverage, their tokenization is not well aligned with the morphological structure of Traditional Mongolian. Therefore, we focus on comparisons with monolingual baselines to better isolate the impact of morphology-aware modeling.

Experiments select Word2Vec(Mikolov et al., 2013) and single-granularity BERT(Devlin et al.,

Language	Model	UAS (%)	LAS (%)
Mongolian	Word2Vec+BiLSTM Parser	79.9	73.6
	BERT	86.1	81.7
	MAGNet-Mongolian (Ours)	**88.7**	**85.0**
Turkish	Word2Vec+BiLSTM Parser	78.6	72.4
	BERT	85.3	80.9
	MAGNet-Turkish (Ours)	**87.9**	**84.1**

Table 1: Dependency Parsing Performance on Mongolian and Turkish

2019) as baseline models, and also incorporate the MAGNet proposed in this paper (with dedicated versions for Mongolian and Turkish) for performance comparison. Downstream tasks cover dependency parsing and named entity recognition (NER) for both languages, where the core task of NER is to identify person, location and organization names in all cases. For ablation studies, three model variants are designed, namely the variant with morphological segmentation removed (–MorphSeg), the variant with morphological tagging removed (–MorphTag), and the variant with the whole-word encoder removed (–Word), so as to verify the functional value of each core component.

4.3 Downstream Task Performance

We evaluate MAGNet on two representative NLP tasks for low-resource agglutinative languages: dependency parsing and named entity recognition (NER). Both tasks are conducted on Mongolian and Turkish respectively. To ensure a fair and controlled comparison, all models share the same task-specific architectures—a BiLSTM-based parser for dependency parsing and a BiLSTM-CRF model for NER—with the input representation as the only varying component. This design isolates the impact of different contextual embeddings and allows us to directly assess their representation quality across tasks and languages.

4.3.1 Mongolian Dependency Parsing and NER

For Mongolian dependency parsing, we report Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS); for NER, we use precision, recall, and F1-score (Table 1 and 2). Word2Vec-based models perform poorly on both tasks (UAS=79.9%, LAS=73.6% for parsing; F1=71.9% for NER), as static word vectors fail to model Mongolian's rich morphological inflections and context-dependent word meanings. Single-granularity Mongolian BERT achieves significant gains (UAS=86.1%, LAS=81.7% ; NER F1=80.0%) by introducing contextualized embeddings, but is limited by whole-word-only modeling that discards morpheme-level morphological information.

MAGNet-Mongolian outperforms all baselines by a consistent margin: it achieves 88.7% UAS and 85.0% LAS (2.6 and 3.3 percentage point gains over Mongolian BERT) for parsing, and 83.4% F1 for NER (3.4 percentage point gain). This improvement stems from MAGNet's multi-granularity design, which integrates morpheme-level morphological features with whole-word-level syntactic and semantic information. For dependency parsing, this integration enhances the model's ability to capture syntactic dependencies shaped by morphological agreement; for NER, it mitigates out-of-vocabulary (OOV) issues by leveraging morpheme composition of rare proper nouns.

4.3.2 Turkish Dependency Parsing and NER

For Turkish dependency parsing, we report Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS) . Similar to Mongolian, traditional Word2Vec-based models perform poorly due to their inability to capture rich morphological variations, achieving only 77.5% UAS and 71.2% LAS.

The monolingual pre-trained model substantially improves performance to 84.7% UAS and 79.3% LAS, demonstrating the effectiveness of contextualized representations. However, as a single-granularity model operating at the word level, it does not explicitly model internal morphological structure, which limits its ability to fully capture the compositional nature of Turkish morphology.

MAGNet-Turkish achieves the best performance, reaching 87.1% UAS and 82.4% LAS, yielding gains of +2.4 UAS and +3.1 LAS over BERT. Notably, the improvement on LAS is larger than that on UAS, indicating that the proposed model is par-

Language	Model	Precision (%)	Recall (%)	F1-Score (%)
Mongolian	Word2Vec+BiLSTM-CRF	72.6	71.2	71.9
	Mongolian BERT	80.8	79.2	80.0
	MAGNet-Mongolian (Ours)	83.8	83.1	83.4
Turkish	Word2Vec+BiLSTM-CRF	73.9	72.7	73.3
	Turkish BERT	81.5	80.3	80.9
	MAGNet-Turkish (Ours)	84.6	83.9	84.2

Table 2: Named Entity Recognition Performance Comparison for Mongolian and Turkish

Language	Model Setting	Precision (%)	Recall (%)	F1-Score (%)
Mongolian	MAGNet-Mongolian	83.8	83.1	83.4
	–MorphTag	82.5	81.6	82.0
	–MorphSeg	81.9	81.0	81.4
	–Word	81.2	80.5	80.8
Turkish	MAGNet-Turkish	84.6	83.9	84.2
	–MorphTag	83.0	82.3	82.6
	–MorphSeg	82.5	81.6	82.0
	–Word	81.7	81.1	81.4

Table 3: Ablation Study Results for Named Entity Recognition (Mongolian & Turkish)

ticularly effective at predicting dependency labels, which often rely on fine-grained morphological cues such as case markers and agreement features.

4.4 Ablation Study

To comprehensively evaluate the contribution of each component in MAGNet, we conduct ablation studies on both Mongolian (dependency parsing and NER) and Turkish (dependency parsing and NER). Using the full MAGNet model as the baseline, we evaluate three ablated variants:

–MorphTag: removes morphological tagging information while retaining morpheme segmentation

–MorphSeg: removes morpheme segmentation entirely, disabling the fine-grained pathway

–Word: removes the whole-word encoder, retaining only the morpheme-level encoder

Results are presented in Table 4 and 3. Across all tasks and languages, removing any component leads to performance degradation, confirming that each component contributes to the overall effectiveness of the model.

Specifically, removing morphological tagging (–MorphTag) results in a moderate performance drop, indicating that explicit functional labels provide useful grammatical signals. Removing morpheme segmentation (–MorphSeg) leads to a more substantial decline, demonstrating that explicit modeling of internal morphological structure is crucial for handling agglutinative languages. Finally, removing the word-level encoder (–Word) causes the largest performance degradation, highlighting the importance of preserving holistic semantic information.

Model Setting	Dependency Parsing (LAS, %)
MAGNet-Mongolian	85.0
–MorphTag	83.4
–MorphSeg	82.8
–Word	82.3
MAGNet-Turkish	84.1
–MorphTag	82.7
–MorphSeg	82.1
–Word	81.5

Table 4: Mongolian Ablation Study Results for Dependency Parsing

Notably, the consistent trends across Mongolian and Turkish suggest that MAGNet effectively captures universal properties of agglutinative languages, where morphological composition and word-level semantics jointly determine meaning.

4.5 Cross-Lingual and Model Generalization Analysis

Across all experiments, MAGNet consistently outperforms single-granularity baselines on both Mongolian and Turkish, two typologically related yet linguistically distinct agglutinative languages. This cross-lingual robustness highlights MAGNet's ability to capture language-agnostic properties of agglutinative morphology, such as morpheme compositionality and inflectional concatenation, through multi-granularity joint modeling—without relying on language-specific rules or handcrafted features.

Looking forward, this generalization capability suggests that MAGNet could be extended to a broader range of agglutinative and morphologically rich languages, such as Kazakh or Uyghur. Moreover, integrating MAGNet with multilingual

or cross-lingual pre-training paradigms may further enhance its adaptability and scalability, paving the way for unified morphological-aware representation learning across low-resource languages.

5 Conclusion

We propose MAGNet, a morphology-aware multi-granularity pre-training framework that explicitly integrates morphological structure into representation learning while dynamically balancing fine-grained and coarse-grained information.MAGNet consists of three core components: a morphology-aware fine-grained pathway that combines morpheme segmentation and tagging to capture morphological regularities, a coarse-grained pathway that preserves holistic word-level semantics, and a gated fusion mechanism that adaptively integrates the two representations based on context. Experiments on two low-resource agglutinative languages—Mongolian and Turkish—demonstrate the framework's effectiveness. Ablation studies confirm that both morphological tagging and whole-word encoding are indispensable, highlighting their complementary roles in capturing grammatical functions and semantic coherence.

Cross-lingual evaluations further validate MAGNet's generalizability, as it consistently outperforms baselines across linguistically distinct agglutinative structures without language-specific engineering. This work provides a viable solution for low-resource agglutinative language modeling and opens new avenues for future research: extending the framework to more morphologically rich languages (e.g., Kazakh, Uyghur) and integrating it with cross-lingual pre-training paradigms to enhance scalability across low-resource language families.

Limitations

Despite the promising experimental results, this study has several limitations, which also point the way to future improvements. First, potential inaccuracies caused by segmentation and annotation tools may affect the experimental outcomes, and the experiments in this study are only conducted on two agglutinative languages (Mongolian and Turkish), without extending the evaluation to a wider range of morphologically rich languages. Subsequent work will supplement experiments on more agglutinative languages (e.g., Kazakh, Uyghur) to further verify the cross-linguistic generalization ability of the proposed framework. Second, the effectiveness of MAGNet has only been validated on two downstream tasks (dependency parsing and named entity recognition), and its performance on other typical natural language processing tasks such as sentiment analysis, machine translation, and part-of-speech tagging has not been explored. Such tasks may reveal additional strengths and weaknesses of the multi-granularity representation learning approach.

Acknowledgments

This study is supported by the 2022 Inner Mongolia Talent Support Pro-ject(Grant No. DC2300001440), the Science Research Foundation of Inner Mongolia Universi-ty of Technology (Grant No. DC2300001262), and the Inner Mongolia Natural Science Foun-dation (Grant Nos. 2024MS06017, 2025MS06036).

References

Ahmet Acar, Onur Güngör, and Zemberek Development Team. Zemberek-nlp: Turkish natural language processing library. GitHub.

Duygu Altinok. 2023. A diverse set of freely available linguistic resources for Turkish. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13739–13750, Toronto, Canada. Association for Computational Linguistics.

Catherine Arnett and Benjamin K. Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 6607–6623. Association for Computational Linguistics.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

M. Creutz and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526, Sofia, Bulgaria. Association for Computational Linguistics.

Bofang Li, Aleksandr Drozd, Tao Liu, and Xiaoyong Du. 2018. Subword-level composition functions for learning word embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 38–48, New Orleans. Association for Computational Linguistics.

Na Liu, Jiajia Ma, Zhonghao Zhang, Aodengbala Aodengbala, Min Lu, and Guiping Liu. 2025. Pre-training and fine-turning multi-task learning model: An effective method for mongolian morphological tagging. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li. 2021. Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:274–285. Doi: 10.1109/TASLP.2020.3040523.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of ICLR*.

Muhan Na, Xiaolin Jin, and Weihua Wang. 2024a. Mongolian Pre-trained Language Model Incorporating Agglutinative Language Features. *Journal of Minzu University of China (Natural Sciences Edition)*, 33(3):32–39.

Muhan Na, Rui Liu, Feilong Bao, and Guanglai Gao. 2024b. Pre-training language model for mongolian with agglutinative linguistic knowledge injection. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of EMNLP*.

M. E. Peters, M. Neumann, M. Iyyer, and 1 others. 2018. Deep contextualized word representations. *Proceedings of NAACL-HLT*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Lao S. and et al. 2016. Mongolian named entity recognition system with rich features. In *Proceedings of COLING 2016*, pages 211–220. 33209PER/LOC/ORG.

Umut Sulubacak and Gülşen Eryiğit. 2018. Implementing universal dependency, morphology and multiword expression annotation standards for turkish language processing. *Turkish Journal of Electrical Engineering & Computer Sciences*, pages 1–23.

Y. Sun, S. Wang, Y. Li, and 1 others. 2020. Ernie: Enhanced representation through knowledge integration. *Proceedings of ACL*.

Changbing Yang and Garrett Nicolai. 2025. Learning beyond limits: Multitask learning and synthetic data for low-resource canonical morpheme segmentation. *ArXiv*, abs/2505.16800.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. Mc^2: Towards transparent and culturally-aware nlp for minority languages in china. *arXiv preprint arXiv:2311.08348*.

Xinsong Zhang, Pengshuai Li, and Hang Li. 2021. Ambert: A pre-trained language model with multi-grained tokenization. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 421–435.