ABSTRACT

Integrating Nanosyntax as Linguistically-Driven Framework in Machine Translation Preprocessing

Gelora Damayanti Manalu
21/480851/PA/20914

Current neural machine translation models inadequately handle spatial prepositions by treating them as probabilistic sequential atomic units, failing to capture their internal compositional structure and resulting in semantic shifts in translations. This study integrates nanosyntax, a linguistic framework that decomposes morphemes into fine-grained syntactic features, as a preprocessing approach for English-Indonesian machine translation. A comprehensive nanosyntax lexicon containing 166 spatial preposition entries was developed based on Starke's nanosyntax and Svenonius's spatial adpositions decomposition theory. The methodology employed constituent tree traversal, followed by nanosyntax lexicalization using phrasal spell-out principles. The translation model combined a GRU encoder-decoder with attention mechanism and Graph Convolutional Networks to process both textual and syntactic tree representations. Experiments were conducted on news corpus, Asian Language Treebank (19807 sentence pairs) and spatial expression datasets (REAL Corpus and ReferItGame). Results demonstrate that the nanosyntax model achieved superior performance with BLEU scores of 26.94 on the news domain test set (+0.64 improvement) and 16.62 on spatial expressions domain test set (+1.35 improvement) compared to baseline. The model exhibited enhanced capability in handling complex sentences and longer texts while providing regularization effects that prevented overfitting. This research validates nanosyntax as an effective linguistically-driven approach for improving machine translation quality, particularly for structurally complex spatial expressions.

Keywords: Computational Linguistic, Natural Language Processing, Machine Translation, Nanosyntax, Preprocessing

ABSTRAK

Integrasi Nanosyntax sebagai Kerangka Berbasis Linguistik dalam Pra-pemrosesan Penerjemah Mesin

Gelora Damayanti Manalu
21/480851/PA/20914

Model terjemahan mesin saraf saat ini tidak memadai dalam menangani preposisi spasial karena memperlakukannya sebagai unit atomik sekuensial probabilistik, sehingga gagal menangkap struktur komposisional internal dan mengakibatkan pergeseran semantik dalam terjemahan. Penelitian ini mengintegrasikan Nanosyntax, kerangka linguistik yang menguraikan morfem menjadi fitur syntax secara sangat rinci, sebagai pendekatan pra pemrosesan untuk terjemahan mesin Inggris-Indonesia. Lexicon nanosyntax komprehensif berisi 166 entri preposisi spasial dikembangkan berdasarkan nanosyntax Starke dan teori decomposition of spatial adpositions Svenonius. Metodologi menggunakan constituent tree traversal, diikuti lexicalization nanosyntax menggunakan prinsip phrasal spell-out. Model terjemahan menggabungkan GRU encoder-decoder dengan attention mechanism dan Graph Convolutional Networks untuk memproses representasi tekstual dan syntactic tree. Eksperimen dilakukan pada korpus berita, Asian Language Treebank (19807 pasangan kalimat) dan dataset ekspresi spasial (REAL Corpus dan ReferItGame). Hasil menunjukkan bahwa model nanosyntax mencapai performa superior dengan BLEU scores 26,94 pada news domain test set (+0,64) dan 16,62 pada spatial expressions domain test set (+1,35) dibandingkan baseline. Model menunjukkan kemampuan yang ditingkatkan dalam menangani kalimat kompleks dan teks yang lebih panjang sambil memberikan efek regularisasi yang mencegah overfitting. Penelitian ini memvalidasi nanosyntax sebagai pendekatan berbasis linguistik yang efektif untuk meningkatkan kualitas terjemahan mesin, khususnya untuk ekspresi spasial yang kompleks secara struktural.

Keywords: Computational Linguistic, Natural Language Processing, Machine Translation, Nanosyntax, Preprocessing