論 文 の 内 容 の 要 旨

論文題目　Interlingual Semantic Analysis of Text:
Alternative methods to full corpus annotation
（テキストの中間言語方式意味的解析：
全コーパスアノテーションの代替法）

氏　　名　ホリエ　アンドレ　ケンジ

In Natural Language Processing, a trend towards shallow linguistic statistical approaches has been observed recently. While these approaches have low implementation costs and present reasonably satisfactory outputs, there is a trade-off on the output quality and naturality observed in such applications. Deep semantic approaches, on the other hand, enable meaning to be better conveyed, despite requiring manual annotation of large corpora to be used by supervised machine learners. These high costs represent one of the main reasons hindering their wider adoption.

This research aims to decrease annotation costs for Interlingual Semantic Computing. It considers semantic domains separately, following a principle in Semantics in which meaning may be broken into its constituents due to its compositional nature, and focuses on analyzing semantic elements that constitute events under a textual context for the selected domains of contextual relations, modality and tense. After analysis on the linguistic properties of the problem, alternative methods which aim to decrease the number of annotated instances, simplify annotation and/or decrease requirements on annotators' level of specialization are presented for each domain.

For the domain of contextual semantic relations, a hybrid bootstrapped set expansion and active learning approach is proposed. It is a semi-supervised process which extracts new instances from an unannotated corpus such as the web. This approach addresses the

problems of class imbalance and incomplete feature spaces, creating feature-rich datasets from initial small seeds and enabling better allocation of annotation resources.

For the domain of modality, cue selection is decoupled from cue expression disambiguation. This allows optimization of selection, which is empirically shown to outperform existing systems in the general case without employing cue expression annotation even when using less resource-intensive disambiguation settings such as lemmatization only.

Finally, for the domain of tense, the analysis task is proposed, addressing previously identified difficulties. The cost of annotation of such approach is then decreased by proposing more intuitive descriptors and by automatically inferring tense, which is only possible because of the proposed novel theory of tense. Unlike other works, which assume extraction of semantic structure as trivial, this theory investigates how tense is perceived and composed from surrounding temporal entities (verbs, adverbials and textual context), which is evaluated and validated through a proof of concept.