論文の内容の要旨

論文題目

Japanese Anaphora Resolution Based on Automatically Acquired World Knowledge

(自動獲得した世界知識に基づく日本語照応解析)

氏名　笹　野　　遼　平

In natural language text, several concepts have tight relations with each other. However, due to the linear constraints of text, most of them are not obvious in the normal form of text; thus automatic recognition of such relations is considered to be an essential step in natural language understanding (NLU). Anaphora resolution, including coreference resolution, zero anaphora resolution and bridging reference resolution, is one of the important subtasks for automatic recognition of such relations.

In this thesis, we focus on Japanese texts. A typical NLU model for Japanese texts first segments input sentences into word sequences, assigns part-of-speech (POS) tags, recognizes named entities (NEs), and then recognizes syntactic structure and case structure. As a consequence of these analyses, relations that are expressed on the surface of text are recognized. In succession to these analyses, anaphora resolution, which resolves relations that are not expressed on the surface of text, are conducted.

The state-of-the-art morpho-syntactic analyzer and NE recognizer are considered to have achieved reasonable performance. However, to recognize more complicated relations, such as coreference relations, more accurate systems are desirable. For example, NEs play an important role in coreference resolution; thus more accurate NE recognition system is considered to benefit the performance of coreference resolution. Therefore, we first aim to improve NE recognition. In Chapter 2, we propose an NE recognition system that uses non-local information. While conventional Japanese NE recognition system has been often performed immediately after morphological analysis and rely only on local context, our system performs after structural analyses and uses four types of non-local information: cache features, coreference relations, syntactic features and case frame features, which are obtained from structural analyses. We evaluated our approach on CRL NE data and obtained a higher F-measure than existing approaches that do not use non-local information. We also conducted experiments on IREX NE data and an NE-annotated web corpus, and confirmed that non-local information improves the performance of NE recognition.

Since there are few grammatical clues for resolving anaphoric relations, world knowledge concerning such relations is necessary to resolve them. For example, synonym knowledge is essential for recognizing

coreference relations between paraphrased mentions; case frames, which describe what kinds of cases each predicate has and what kinds of nouns can fill these case slots, are essential for zero anaphora resolution. There have been some studies that have tried to elaborate these knowledge by hand, but the problem is their coverage. That is to say, it is very difficult to make wide-coverage knowledge manually, because language is composed of an enormous number of content words. Moreover, there are technical terms or jargon for every domain, and new words are coined every day. In Chapter 3, we describe how to acquire world knowledge automatically. We first extract synonym knowledge, which is utilized for coreference resolution, from a large raw corpus and dictionary definition sentences. Secondly, we construct case frames from modifier-head examples in the resulting parses of large corpora. The problems for case frame construction are syntactic and semantic ambiguities. To cope with these problems, the case frames were gradually constructed from reliable modifier-head examples. Furthermore, in order to deal with data sparseness problem, we generalize the examples of case slots. Finally, we construct nominal case frames, which describes indispensable entities of nouns and utilized for bridging reference resolution. The point of the construction method is the integrated use of a dictionary and example phrases from large corpora.

Chapter 4 presents a knowledge-rich approach to Japanese coreference resolution. In Japanese, proper noun coreference and common noun coreference occupy a central position in coreference relations. To improve coreference resolution for such language, wide-coverage synonym knowledge is utilized. Furthermore, to boost the performance of coreference resolution, we integrate primitive bridging reference resolution system into coreference resolver. The experimental results show that using the acquired synonyms and bridging reference resolution boosted the performance of coreference resolution and the effectiveness of our integrated method is confirmed.

Chapter 5 presents a probabilistic model for Japanese zero anaphora resolution. First, this model conducts coreference resolution, recognizes discourse entities and links all mentions to them. Zero pronouns are then detected by case structure analysis based on automatically constructed case frames. Their appropriate antecedents are selected from the entities with high salience scores, based on the case frames and several preferences on the relation between a zero pronoun and an antecedent. Case structure and zero anaphora relation are simultaneously determined based on probabilistic evaluation metrics.

Chapter 6 reports the effect of corpus size on case frame acquisition for discourse analysis. For this study, case frames were constructed from corpora of six different sizes ranging from 1.6 million to 1.6 billion sentences. These case frames were then applied to syntactic and case structure analysis, and zero anaphora resolution. Better results were obtained by using case frames constructed from larger corpora; the performance was not saturated even with a corpus size of 1.6 billion sentences.

Chapter 7 provides concluding remarks, summaries the thesis, and outlines the areas for future work.