氏　　　名	JOHN　PHILIP MCCRAE		
学位（専攻分野）	博士（情報学）		
学位記番号	総研大 1288 号		
学位授与の日付	平成２１年９月３０日		
学位授与の要件	複合科学研究科　情報学専攻 学位規則第６条第１項該当		
学位論文題目	Automatic extraction of logically consistent ontologies from text corpora		
論文審査委員	主　査	准教授	Nigel Collier
		教授	井上　克巳
		教授	武田　英明
		教授	佐藤　健
		准教授	古山　宣洋
		准教授	Hong YU (U. of Wisconsin-Milwaukee)

論文内容の要旨

Automatic extraction of logically consistent ontologies from text corpora

Ontologies provide a structured description of the concepts and terminology used in a particular domain and provide valuable knowledge for a range of natural language processing applications. However, for many domains and languages ontologies do not exist and manual creation is a difficult and resource-intensive process. As such, automatic methods to extract, expand or aid the construction of these resources is of significant interest.

There are a number of methods for extracting semantic information about how terms are related from raw text, most notably the approach of Hearst [1992], who used patterns to extract hyponym information. This method was manual and it is not clear how to automatically generate patterns, which are specific to a given relationship and domain. I present a novel method for developing patterns based on the use of alignments between patterns. Alignment works well as it is closely related to the concept of a join-set of patterns, which minimally generalize over-fitting patterns. I show that join-sets can be viewed as an reduction on the search space of patterns, while resulting in no loss of accuracy. I then show the results can be combined by a support vector machine to a obtain a classifier, which can decide if a pair of terms are related. I applied this to several data sets and conclude that this method produces a precise result, with reasonable recall.

The system I developed, like many semantic relation systems, produces only a binary decision of whether a term pair is related. Ontologies have a structure, that limits the forms of networks they represent. As the relation extraction is generally noisy and incomplete, it is unlikely that the extracted relations will match the structure of the ontology. As such I represent the structure of ontology as a set of logical statements, and form a consistent ontology by finding the network closest to the relation extraction system's output, which is consistent with these restrictions. This gives a novel NP-hard optimization problem, for which I develop several algorithms. I present simple greedy approaches, and branch and bound approaches, which my results show are not sufficient for this problem. I then use resolution to show how this problem can be stated as an integer programming problem, which can be efficiently solved by relaxing it to a linear programming problem. I show that this result can efficiently solve the problem, and furthermore when applied to the result of the relation extraction system, this improves the quality of the extraction as well as converting it to an ontological structure.

博士論文の審査結果の要旨

Automatic extraction of logically consistent ontologies from text corpora

The thesis is structured into 6 sections which are summarized below:
1) Introduction. Ontologies as conceptual representations for a domain of knowledge are introduced and their contribution to the field of natural language processing (NLP). A distinction is made between generic resources such as WordNets which are widely used but weak in domain vocabulary and domain-specific ontologies which are potentially powerful but scarce and expensive to construct. The idea is then introduced for automatically extending WordNet style ontologies using terminology and relationships discovered in domain text collections. Logical restrictions in the relationships require explicit modeling but developing an algorithm that can integrate such restrictions is NP-complete. The aim of the thesis is stated as to discover consistent ontologies from text by (a) extracting relationships between domain based terminology automatically from texts, and (b) integrating the discovered binary relationships between terms into a structured ontology whilst respecting the logical restrictions.
2) Background survey of related work. Three main sections are given: (1) WordNets, the connection between logic and ontologies, and ontology applications in natural language processing and the Semantic Web communities; (2) Related work on extracting semantic relations is divided into three main classes based on source of knowledge, the co-occurrence of related terms in a single context, usually extracted by patterns, the extraction of semantic knowledge by the contexts terms occur in independently and methods for clustering this information, and the similarities of the surface forms of the terms. A comparison between the strengths and weakness of each approach is presented. (3) Review of work on forming ontological structures including hierarchical clustering and Snow et al's approach based on logical restrictions. The key issue that emerges is the challenge of forming a consistent ontological structure from incomplete and noisy snippets of data found in the corpus. The approach adopted in the remainder of the thesis handles this by efficiently incorporating logical axioms into the learning process to describe the expected structure of the ontology whilst dealing with the NP-hard nature of the resulting algorithm.
3) Relation extraction from corpora. A novel pattern generation language and algorithm is developed based on literals, entity types and wildcards which is then constrained using pseudo accuracy measures that can extract multi-word terms. Join sets are then developed to efficiently reduce the size of the search space from a set of rules discovered in corpora. The section finished by introducing the idea of classification for finding the probability that a given term pair will be related by a specific relationship.
4) Logical inconsistencies of extracted ontologies. The notion of a network of relationships between a set of elements is introduced. A cost function is developed for finding a valid subnetwork that under a reasonable independence assumption leads to a linear function based on the probability values from the pattern-based extraction system. First the problem of finding the optimal network over a set of mutually exclusive set – i.e. to find the exact cover of the elements and an algorithm is presented for this. Then a definition is given for more general logic restrictions and using the theorem proving method called resolution this is then converted into an integer programming problem which is efficiently solved

by relaxing it to a linear programming problem. This allows the problem to be solved by using the simplex algorithm to solve the linear programming problem and then using a branch-and-bound approach to solve the remaining integer programming problem. A weighted MAX-SAT problem is proposed for dealing with different kinds of network structures based on the use of axioms and an adapted GSAT method is shown which is practical for most situations. Finally it is shown that extant OWL (Web Ontology Language) ontologies can be handled within this framework.

5) Results. Based on a human gold standard of synonym term sets in 150 PubMed abstracts experiments are performed which show how the new synonym pattern learner based on (3) compares to human standards for extracting synonym relations in terms of recall, precision and F-measure. The new learner (F-measure 35.7) is shown to outperform Wikipedia, WordNet, Medline Encyclopedia and MeSH but not UMLS. Since UMLS is the product of many millions of man hours of work it is concluded that the automatic method has considerable merits in terms of accuracy and throughput. The top set of discovered patterns is reported and compared to Hearst's manually built patterns. Using synonym pairs from the BioCaster ontology further experiments show how various standard machine learning algorithms (e.g. SVM, Logistic regression) used with logical consistency checking could be used to group discovered terms into synsets. Further experiments show the utility of the join set methodology by incorporating restrictions on hyponymy and synonymy using the subset of 'disease' in WordNet as the gold standard and PubMed documents as the test collection. Finally simulated data is used to show the algorithmic complexity of the improved method shown in (4) against 6 baselines including GSAT.

6) Conclusion. The novel contribution of the thesis is summarized from several points: (a) The new method for ontology extraction from text is comparable and in many cases better than existing human resources for the real world domain and examples covered; (b) Issues such as the explosion in the number of patterns arising from specialization of successful patterns were successfully dealt with by exploiting the connection between the join–set concept and the alignment of two patterns; (c) The effectiveness of the algorithms was shown for obtaining logical consistency and obtaining an optimal solution in a short amount of time.