Expert Systems with Applications 41 (2014) 7171–7181

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Multi level causal relation identification using extended features

Xuefeng Yang *, Kezhi Mao

School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, 639798, Singapore

ARTICLE INFO

Article history:
Available online 9 June 2014

Keywords:
Machine learning
Knowledge acquisition
Text mining
Information extraction

ABSTRACT

Extracting causal relation underlying natural language is an important issue in knowledge discovery. Most previous studies of casual relation extraction focus on simple cases like causal relations between two noun phrases indicated by fixed verbs or prepositions. For more complicated causal relations, such as causal relations between clauses, the previously developed algorithm may not work. To solve this problem, this paper develops a system that is able to extract causal relations in multi-level language expressions such as, words, phrases and clauses without fixed relators. The information extraction system is composed of a multi-level relation extractor and an ensemble-based relation classifier. It may extract more subtypes of causal relations than previous work because extracting domain is expanded in terms of both syntactic expressions and semantic meanings. In addition, the proposed method outperforms previously developed methods because extended features based on lexical semantic resources are explored. Experiments show that our system achieves an accuracy of 88.69% and *F*-score of 0.6637 in a dataset with 300 sentences.

1. Introduction

Natural language is the major intermediary tool for human communication. However, natural language text is unstructured and hard to understand for computers. Transferring the unstructured language text into a machine readable model has been researched from distinct perspectives, for example, the Message Understanding Conferences (Grishman & Sundheim, 1996) and Automatic Content Extraction (ACE) programme (Doddington et al., 2004). Most studies focus on extraction of descriptive information like entities, relations and events, and less attention has been paid for logic level information extraction, though it may be more meaningful. Causal relation is the most important logic relation in natural language, helping connect various descriptive information into meaningful sentences and paraphrases. The importance of causal relation may be illustrated briefly by the following example: *Jim is happy today since his favorite basketball team won the final*. Most relation extraction systems focus only on relations between two words like *Be*(Jim, Happy) and *Win*(Team, Final), however the relations *Be*(Jim, Happy) and *Win*(Team, Final) is less meaningful without the causal relation (Jim is happy)*Since* (His favorite basketball team won the final).

Causal relation identification is useful in a variety of natural language processing tasks. Khoo, Myaeng, and Oddy (2001) employed cause-effect pairs to improve information retrieval. Girju (2003) used causal relation in unstructured text to enhance Question and Answering system. Bayesian network and causal association network are popular for decision support applications, and Sanchez-Graillet and Poesio (2004), Raghuram et al. (2011) constructed the networks from unstructured text using causal relations between concepts. Causal relation extraction also benefits biomedical knowledge summarization. Khoo, Chan, and Niu (2000) employed graphical pattern to extract causal relation in medical textual database, and Bui, Nualláin, Boucher, and Sloot (2010) used a series of natural language processing tools to retrieve relations between drugs and virus mutations from medical literatures.

Despite of its importance, causal relation extraction is very challenging due to complicated syntactic expressions and numerous fine-grained sub-concepts. Causal relation may be categorized as inter-sentential or intra-sentential. Inter-sentential causal relation extraction aims to find causal relations between events at document level (for example Do, Chan, & Roth (2011) and Radinsky, Davidovich, & Markovitch (2012)). This paper focuses on intra-sentential causal relation, whose cause and effect might be words, phrases or clauses from the same sentence.

Some early studies (see examples in Joskowicz, Ksiezyck, & Grishman (1989) and Kaplan & Berry-Rogghe (1991)) attempted

* Corresponding author. Tel.: +65 97920819.
E-mail addresses: yang0302@e.ntu.edu.sg (X. Yang), EKZMao@ntu.edu.sg (K. Mao).

http://dx.doi.org/10.1016/j.eswa.2014.05.044
0957-4174/

to extract causal relations based on inference with predefined knowledge bases, while others used linguistic patterns without domain knowledge (for example Khoo et al. (2000)). In recent years, machine learning has been employed for causal relation extraction because machine learning-based approach is more robust than pure rules based approaches, and it requires less linguistic and domain knowledge. There are three main steps in machine learning-based approach, including candidate relation extraction, feature generation and classification. Initially, all possible relations underlying a sentence are extracted as candidate relations. The extracted candidate relations are then encoded into the numerical representation in the feature generation step. Finally, the encoded candidate relations are classified as causal or non-causal in the classification step.

Girju and Moldovan (2002) extracted < *NP1 causal-verb NP2* > syntactic patterns with causative verbs and then employed semantic constraints to classify candidates as causal or non-causal. This approach firstly generate regular patterns based on expert knowledge to search for the matched patters in the internet and large collection of documents, and then impose some constrains for *NP1, NP2* and verbs to filter out the non-causal candidates. (Girju, 2003) modified the previous work and used C4.5 decision tree instead of simple constraints to perform classification for Question and Answering application. The causal relation is restricted to noun pairs with a verbal expression as the connection word. Chang and Choi (2006) used relator phrase and words pairs to generate candidates from large corpus and then employed probabilistic model to classify the patterns based on the probability of relator phrase and noun words pair. The candidates are generated by selecting predefined patters from dependency parsed tree structure. The cause and effect are noun phrase and the connection words are the same as those in Girju and Moldovan (2002). Beamer, Rozovskaya, and Girju (2008) proposed a novel boundary feature extracted from WordNet to help causal relation classification between nominal. In Blanco, Castell, and Moldovan (2008), the authors employed predefined syntactic patterns to extract candidates containing any of the following four relators: "because", "after", "as" and "since", and then classified the patterns using a bagging ensemble of tree classifiers.

Although intra-sentential causal relations exist in many different language expression levels, previous studies may only extract causal relations satisfying the predefined patterns such as relations between two noun phrases, relations indicated by preposition and relations containing selected verbs. For the causal relations in the clauses level, for instance, the given example (Jim is happy)*Since* (His favorite basketball team won the final) which is very common in natural language expressions, the algorithm developed in previous studies may not work. To fill the gap, the authors develop a multi-level causal relation extraction algorithm based on the linguistic knowledge of the dependency grammar and constituent grammar without imposing any restriction on causal relation patterns.

Removing the pattern restrictions means that the domain of causal relations is expanded in both syntactic and semantic perspectives. However, the larger coverage of causal relation expressions causes feature deficiency problem which means that the features used in previous studies may not perform well in identifying complex explicitly expressed intra-sentential causal relations. The reason behind this is the information insufficiency of the employed features. This motivates us to explore more representable features for this more challenging task. Employing knowledge stored in different resource is the most widely used solution to address the feature deficiency problem in applications such as text classification (Meng, Lin, & Li, 2011), text clustering (Jing, Ng, & Huang, 2010) and information retrieval (Varelas, Voutsakis, Raftopoulou, Petrakis, & Milios, 2005). The linguistic knowledge included in lexical semantic resource is extracted by the rules defined in this paper, most of which are not used in previous studies. Three lexical semantic resources are used in our study, including WordNet (Fellbaum, 2010), VerbNet (Schuler, 2005) and FrameNet (Baker, Fillmore, & Lowe, 1998). WordNet is widely employed for semantic research (Budanitsky & Hirst, 2006; Li, Yang, & Park, 2012; Lee, Huh, & McNiel, 2008) and has been proven useful for causal relation classification. VerbNet and FrameNet have not been employed for causal relation classification task in previous studies.

However, because of curse of dimensionality, the expansion of features to represent relationship does not mean improved performance. To obtain a compact feature subset consisting of features with good separation capability, feature selection techniques are employed to rank the features based on the data.

An merit of the proposed system is the use of both domain knowledge and data in feature generation and relation classification. Although data driven classification techniques are good at quantitatively choosing optimal parameters to approximate the data distribution, they may not perform well with just raw string input representation. Quite often, a single rule based on expert knowledge can extract high level representable features and significantly reduce the classification complexity. In the developed system, expert knowledge rules in the bottom layers of the system to process sentence and generate numerical representation, and data driven classifiers in the top layer makes the final decisions.

The remainder of the paper is organized as follows. Section 2 introduces the multi-level relation extractor for this application. Feature exploration and selection are detailed in Section 3. Ensemble classifier and Experiments are given in Section 4 and 5 respectively. Section 6 concludes the paper.

2. Multi-level relation extractor

Previous studies focus on subsets of syntactic patterns for causal expressions, which may not cover causal relations in multi-level language expressions. To extract these complicated causal relations, all explicitly expressed relations in a sentence should be extracted as candidates first. For the relation extraction task, *OpenIE* systems such as Reverb (Etzioni, Fader, Christensen, Soderland, & Mausam, 2011) focus on relations between noun phrases. They target web-scale documents and extract millions of tuples to generate databases (Etzioni, Banko, Soderland, & Weld, 2008; Etzioni et al., 2011), but the output reserves little information for further analysis. *OBIE* system only extracts predefined relations in the ontology (see for example Wimalasuriya & Dou (2010) and Moreno, Isern, & LóPez Fuentes (2013) and the references therein). These systems pursue large-scale, speed and the most important relations in text, but may not be suitable for deep semantic analysis of relations in a sentence where every relation may be important. In addition, all these works use tuple format *Relation*(Arguments) to indicate relations, where the *Arguments* cannot be any other relations.

In this study, a multi-level relation extractor (*MLRE*)[1] is developed which may extract almost all potential relations with any verb or preposition. The relation extractor is built on Stanford Parser (Klein & Manning, 2003), which includes two major parsed relation formats, namely the dependency tree and constituent tree. The output of Stanford Parser only indicates relations between two words which may not satisfy the requirement of

[1] The implementation code for relation extractor is available in https://github.com/YangXuefeng/MLRE.