Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities

Du-Seong Chang* and Key-Sun Choi

Department of Electrical Engineering & Computer Science, KORTERM, BOLA
Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejeon, 306-701, Korea
dschang@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Abstract. This work aims to extract causal relations that exist between two events expressed by noun phrases or sentences. The previous works for the causality made use of causal patterns such as causal verbs. We concentrate on the information obtained from other causal event pairs. If two event pairs share some lexical pairs and one of them is revealed to be causally related, the causal probability of another event pair tends to increase. We introduce the lexical pair probability and the cue phrase probability. These probabilities are learned from raw corpus in unsupervised manner. With these probabilities and the Naive Bayes classifier, we try to resolve the causal relation extraction problem. Our inter-NP causal relation extraction shows the precision of 81.29%, that is 7.05% improvement over the baseline model. The proposed models are also applied to inter-sentence causal relation extraction.

1 Introduction

Causality or *causal relation* refers to 'the relation between a cause and its effect or between regularly correlated events'[1]. Even if not a few questions order to find causality from text, the current Question Answering system cannot respond causal questions. The recent Question-Answering system can produce correct answers to 83.0% of questions (Moldovan et al., [2002]). But the answer accuracy has a wide variation across the question type. Moldovan et al. ([2003]) states the relatively high answer accuracy on questions about the person, the time, the location, and so on. They show very low performance on causal questions. Causal questions are answered with a low precision score of 3.1%. Since there are few causal questions in their test suite of TREC(Text REtrieval Conference), total performance is high in spite of the low performance on causal questions. However, causal questions are very frequently used in an actual question answering. For a web site[2] in which users exchanged questions and answers, there are

* This research was supported in part by KISTEP Strategic National R&D Program under contract M1-0107-00-0018 and by KOSEF under contract R21-2003-000-10042-0.

[1] From Merriam-Webster's Online Dictionary.

[2] Naver Knowledge iN, http://kin.naver.com

K.-Y. Su et al. (Eds.): IJCNLP 2004, LNAI 3248, pp. 61–70, 2005.

130,000 causal questions from 950,000 sentence-sized DB. This fact shows that it is necessary to analyze the causal relation for a high performance Question-Answering.

To response the causal question such as (1a), the following problems should be solved. The first problem is *event extraction*, which extracts events from the paragraphs including keyword 'hiccups' of the question. *Event* is defined as 'a phenomenon or occurrence located at a single point in space-time'[3]. The second problem is *causal relation extraction*, which analyzes the causal relation between events. *Causal question answering* is the last one to be solved. It infers the answer to the question. In this paper, we concentrate the causal relation extraction.

(1a) What are hiccups caused by?
(1b) The oral bacteria that cause gum disease appear to be the culprit.

Cue phrase is a word, a phrase, or a word pattern, which connects one event to the other with some relation. The causal relation between events is assumed by the cue phrase. The causal cue phrase is used for connecting the cause and effect events. When events are expressed by noun phrases, the cue phrase connecting events is a verb phrase in general. For example, in (1b), the verb 'cause' is a cue phrase to connect two events expressed by noun phrases, 'the oral bacteria' and 'gum disease'. Several lexical pairs are assumed to lead the causal relation. The lexical pair 'bacteria' and 'disease' is an example of the causal lexical pair. If the term pair 'the oral bacteria' and 'gum disease' is causally related, we can infer that the event pair 'bowel bacteria' and 'bowel disease' is causally related. Causal lexical pairs are learned from cause-effect event pairs. We define *lexical pair probability* as the probability of the lexical pair that is a part of causal event pairs. The pair of concept classes of each event, [B03] and [C23.550.288][4], also lead the causality of event pair. *Conceptual pair probability* is defined as the probability of the conceptual pair that has the causal relation. Cue phrases connecting two events are also considered to have connection probability. We define *cue phrase probability* as the probability of the cue phrase that connects causal event pairs. With these probabilities, we introduce a causal relation classifier based on Naive Bayes classifier. These probabilities are learned from the raw corpus in an unsupervised manner.

In section 2, selected works are compared for the causal relation extraction. Our classification model will be explained in section 3. In this paper, we aim to extract the causal relation that exists between two noun phrases or sentences. In section 4, we evaluate the proposed model for the inter-noun phrase causality extraction, and prove it adaptable to the inter-sentence causality extraction.

[3] From The American Heritage Dictionary of the English Language: Fourth Edition, 2000.

[4] They represent [bacteria] and [disease]. These conceptual numbers follow the biomedical ontology (Medical Subject Heading, [2004]).