Project notes on building a conversation parser on top of a text parser: Towards a causal language tagger for spoken Chinese

Andreas Liesenfeld
Nanyang Technological University, Singapore
Heinrich-Heine Universitaet Duesseldorf, Germany
lies0002@ntu.edu.sg

Abstract

This ongoing doctoral study examines cause and effect relationships in Chinese spoken language corpora and aims to build a tagger (Cause-Chi) that automatically annotates linguistic patterns used to express these relationships. Drawing on insights from Construction Grammar (CxG), Cause-Chi is a tool to detect explicit causal language and automatically parse constructions of causation and their slot-fillers for Chinese conversational corpus data. Built on top of an existing tagger for text corpora, Cause-Chi is designed to not only detect lexical constructions but also conversation-specific causal language such as multi-segment causal expressions and the usage of temporal constructions to express causal relation. Cause-Chi is currently under development and will be released in 2018 together with MYCanCor, a small corpus of spoken Chinese, and a mini-constructicon of causal constructions based on the corpus.

1 Project outline

Cause-Chi is a causal language tagger for spoken Chinese born out of the idea to build a system that can extend shallow semantic parsing beyond lexical triggers on the sentence level. Even when limiting the scope of the project to explicitly stated relationships of cause-effect, this has proven to be a particularly difficult task considering the wide variety of linguistic and other behavioral patterns that we use to express cause-effect relationships in conversation (Wolff et al., 2005).

This paper focuses on several issues that were encountered in the process of building a conversation parser on top of a text parser. Most available CxG-based semantic parsers are designed for text data, not conversational data (e.g. most FrameNet parsers such as SEMAFOR and among others Das et al., 2014; Roth and Lapata, 2015; Taeckstroem et al., 2015). Since the semantic annotation and parsing community already has mature, well-studied tools for parsing expressions of causal relationships in texts, a common approach is to build on top of an existing tool for text data analysis instead of building a new spoken language parser from scratch. This from-text-to-speech approach might be of interest to the dialog annotation community because it describes how existing text parsing systems can be extended to conversation data. In the remainder of this paper, we outline some of the problems (and solutions) that were encountered in the process of rebuilding the annotation scheme of a text tagger to take conversational data as input.

2 Annotating Chinese lexical constructions of causality

The core of the existing system is the text tagger ”Causeway” for lexical constructions of causality designed on New York Times corpus data (Dunietz et al., 2015; 2017). This tagger uses supervised approaches to learn causal relationships and to identify slot-fillers of the constructions in the corpus based on a list of around 170 lexical cues of constructions of causal language (Dunietz et al., 2015). It is only concerned with explicit causal language within the boundaries of a sentence. The tagger only

annotates complete tuples consisting of three elements: a **causal connective**, a **cause** phrase or clause expressing an event or state, and an **effect** event or state.
Cause-Chi not only redesigned this existing system for lexical cues of Chinese (not English) but also extended it to take functional segments in dialog acts as input. To this end, the existing annotation framework was converted to align with the ISO 24617-8 core annotation scheme, following the mapping guidelines laid out in Bunt et al. (2016). ISO DR-Core was chosen because it offers a theory-neural standard to map semantic information that can be integrated with CxG frameworks. Cause-Chi tags causal connectives in functional segments using the DiAML markup language in order to identify causal relationships (”**ISO DRel 1: Cause**”). This relation can be stated using a large variety of possible expressions of causality including verbs, conjunctions, adjectives and MWEs (see Table 1).

Type of construction	**Example (from the MYCanCor corpus)**
Verb	ling6 (’make, to let’), gaau2 dou3 (’make, cause’)
Conjunction	so2 ji5 (’so, therefore), jan1 wai6 (’because’)
Adjective	zung6 jiu3 (’important’), jim4 zung6 (’serious, grave’)
MWE	bat1 dak1 liu5 (’extremely’)
Nominal	git3 gwo2 (’result’), hau6 gwo2 (’consequence’)
Preposition	bing6 (’furthermore’), dou3 (’to, until’)
Complex	jyut6 loi4 jyut6 (’more and more’
Other	jyun4 (’completion marker’), gam3 (’so’)
(Temporal)	jin4 hau6 (’after’), gan1 zyu6 (’and then’)

Table 1: Examples of explicit constructions of causal language from the MYCanCor Spoken Chinese corpus in Jyutping romanization for Cantonese.

Preliminary tests show that Cause-Chi achieves relatively good results for tuples (causal connective, cause, effect) that are located within one turn. As soon as this is not the case, however, the detection rate drops dramatically. These results are in line with previous studies that have shown that connective detection in speech is a more challenging task than in text (Riccardi et al., 2016).
We aim to incorporate additional ways of how causal relationships are expressed in conversation by gradually taking more complex and multi-segment constructions of causality into account. Two phenomena will be described here. Manual annotation of causal language in the MYCanCor corpus of spoken Chinese revealed two frequently used patterns of expressing causality that the above lexical-item based tagging approach does not capture: **(1) the establishment of causal relationships using more than one functional segment in more than one turn** and **(2) the usage of temporal constructions to express causal relationships**.

3 Multi-segment constructions of causality

Multi-segment causal expressions refer to cause-connective-effect tuples that are either interrupted by dialog acts or span over more than two turns. These causal relations are not identified by the vanilla tagger because the cause-effect elements are not directly preceding or following the causal connective. In Example (1) the stated expressed causal relation is ”I don’t join, so I don’t go.” using the connective ”jan1 wai6” (because). Here, only the cause ”don’t join” is in proximity to the connective. The annotation of the full connective-cause-effect tuple fails because a number of dialog acts precede the effect. This causes the tagger to dismiss the incomplete tuple instead of correctly identifying the effect ”not going”. The problem can in some cases be solved by increasing the parsing range to detect a complete tuple or by correctly annotating the dimension and coherence relation of the functional elements in question (Example 1.1).

(1) P1: Yinwei wo bu canjia. Zhidao ma.
P2: Wo zai ting.

P1: Jiu bu qu le.
(This is a Chinese example in Pinyin romanization from the MYCanCor corpus)

’P1: Because I don’t participate. You Know.’
’P2: I’m listening.’
’P1: I’m just not going.

(1.1) DiAML Representation:

```
 <dialogAct id="a1" target="#s1" sender="#p1"
 addressee="#p2"\" dimension="task"
  communicativeFunction="inform" />
 <dialogAct id="a2" target="#s2" sender="#p2"
addressee="#p1" dimension="autoFeedback" />
  <dialogAct id="a1" target="#s1" sender="#p1"
 addressee="#p2"\" dimension="task"
communicativeFunction="inform" /> <dRel xml:id="r1"
  target="#s2" rel="cause"/> <drArg xml:id="e2" target="#s4" />
  <explDRLink rel="#r1" result="#da1"reason="#e2"/>
```

4 The use of temporal language to express causality

Manual annotation of the MYCanCor corpus has shown that temporal constructions such as the completion marker ”wan” (after, finished) are frequently used to express causal relationships in conversation. Since temporal constructions are not identified by Cause-Chi as causal connectives, the tagger does not identify a causal relationship in these cases. Correct tagging of the use of temporal language to express causality has proven to be a rather challenging task because the tagger has to somehow disambiguate between the ”regular” use of temporal constructions and the use of these constructions for the purpose of causality. In Example (2), P2 uses the ”wan” construction to express a causal relationship between ”seeing the doctor” and ”feeling much better”. Cause-Chi does not identify this causal relationship because no causal connective is used.
A possible way to solve this problem is to test whether the used temporal (or completion) construction can be replaced by a causal construction (such as ”because”) by learning possible slot-fillers in the preceding and following functional segments. This way, P2’s dialog act ”After I saw the doctor, I feel much better” could be restated as ”Because I saw the doctor, I feel much better”, given that the ”seeing a doctor” and ”feeling better” constructions appear often enough with the ”because” connective in the corpus. Alternatively, causal relationships could also be inferred between events (or states) by learning pairs of related predicates (”see doctor” - ”feel better”) (Hu et al. 2017). These methods, however, deal with implicit causal relationships and go beyond the scope of this project in its current state.

(2) P1: Ni jintian zenmeyang?
P2: Kan wan yisheng ganjue hao duo le.
(This is a Chinese example in Pinyin romanization from the MYCanCor corpus)

’P1: How are you today?’
’P2: After I saw the doctor, I feel much better.’

(2.1) DiAML Representation:

```
 <dialogAct id="a1" target="#s1" sender="#p1"
```

```
  addressee="#p2"\" dimension="task"
    communicativeFunction="inform" />
  <dialogAct id="a2" target="#s2" sender="#p2"
addressee="#p1" dimension="task" />
communicativeFunction="inform" /> <dRel xml:id="r1"
    target="#s2" rel="cause"/> <drArg xml:id="e2" target="#s4" />
    <explDRLink rel="#r1" result="#da1"reason="#e2"/>
```

5 Discussion and future work

An enhanced version of Cause-Chi that captures the above stated two phenomena is currently in development. Cause-Chi's modular design allows to add features that are able to capture additional ways of how causal relationships are expressed in Chinese, building on the core lexical-trigger parser. Two of such modules are designed to deal with the above mentioned multi-segment and temporal constructions of causality. Possible additional modules could deal with the distinction between degrees and types of causation, addressing a shortcoming of the current tagger that threats causality as binary.
Cause-Chi is expected to be released in 2018 together with MYCanCor, a small corpus of spoken Chinese, and a mini-constructicon of causal constructions based on the corpus.

References

[1] BUNT, Harry; PRASAD, Rashmi. ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In: Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12). 2016. p. 45-54.

[2] DAS, Dipanjan, et al. Frame-semantic parsing. Computational linguistics, 2014, 40.1: 9-56.

[3] DUNIETZ, Jesse; LEVIN, Lori S.; CARBONELL, Jaime G. Annotating Causal Language Using Corpus Lexicography of Constructions. In: LAW@ NAACL-HLT. 2015. p. 188-196.

[4] DUNIETZ, Jesse; LEVIN, Lori; CARBONELL, Jaime. Automatically Tagging Constructions of Causation and Their Slot-Fillers. Transactions of the Association for Computational Linguistics, 2017, 5: 117-133.

[5] HU, Zhichao; WALKER, Marilyn A. Inferring Narrative Causality between Event Pairs in Films. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbruecken. 2017.

[6] RICCARDI, Giuseppe; STEPANOV, Evgeny A.; CHOWDHURY, Shammur Absar. Discourse connective detection in spoken conversations. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016. p. 6095-6099.

[7] ROTH, Michael; LAPATA, Mirella. Context-aware frame-semantic role labeling. Transactions of the Association for Computational Linguistics, 2015, 3: 449-460.

[8] TAECKSTROEM, Oscar; GANCHEV, Kuzman; DAS, Dipanjan. Efficient inference and structured learning for semantic role labeling. Transactions of the Association for Computational Linguistics, 2015, 3: 29-41.

[9] WOLFF, Phillip, et al. Expressing Causation in English and Other Languages. 2005.