

# 基于深度学习的因果知识抽取系统

**V1.0**

# 使用说明书

中山大学-版权所有

## 目录

1 引言 .....	3
1.1 编写目的 .....	3
1.2 术语和缩略词.....	4
2 软件概述.....	4
2.1 软件功能 .....	4
2.2 算法.....	5
2.3 软件运行 .....	7
2.1 系统要求 .....	7
3 系统使用.....	8
3.1 系统启动 .....	8
3.2 输入管理 .....	9
3.3 序列标注结果生成.....	10
3.4 因果知识分析.....	12

# 1 引言

## 1.1 编写目的

近年来，随着人工智能技术的不断发展，语义关系的自动抽取对于许多自然语言处理任务（例如：问题回答、信息检索、事件预测及文本生成等）变得日益重要。语义关系（例如：条件关系、相关关系及因果关系等）用以表示不同事件或实体间是如何相互作用的，其中因果关系由于可以用于事实推理进而影响决策制定，所以在人类认知世界的过程中扮演了十分重要的角色。对事件或实体间的因果关系进行抽取，可以了解信息之间的来龙去脉，获取信息的演化关系，有助于预测和决策。这种因果知识发现在很多领域（例如：金融、医学、生物学、环境科学等）是非常有价值的。

目前主流的因果关系抽取方法可以分为基于规则的方法，基于模式匹配与统计概率和机器学习相结合的方法。仅依赖规则进行模式匹配的方法往往通用性不强，解决特定领域问题时可能会需要大量领域知识，且制定规则耗费大量时间和人力；基于模式匹配与统计概率和机器学习相结合的方法又往往需要大量的特征工程，严重依赖于人工选择文本特征，需要耗费大量时间与精力，且人工选择的特征一般比较简单，难以捕捉上下文的深层语义信息。而依靠强大的表示学习能力，与浅层机器学习方法相比，深层神经网络可以自动发掘特征，极大地减少了特征工程，节约了人力和时间。以长短时记忆网络 LSTM（Long Short-Term Memory）为例，这种循环神经网络由于可以充分

考虑文本的时序信息与上下文的深层语义信息，现已被广泛应用于自然语言处理任务，并在许多序列标注任务（例如：词性标注、命名实体识别、组块化）中取得了 **state-of-the-art** 的效果，LSTM 对句子语义建模的能力是毋庸置疑的。

本系统就是将因果知识抽取转化为序列标注问题，并结合深度学习的方法和技术，以最大限度地减少特征工程，有效地发现并捕捉关于因果关系的语义表征，对文本中因果关系有效建模。

## 1.2 术语和缩略词

Bi-LSTM 双向长短时记忆网络

CRF 条件随机场

Softmax 一种用于多分类的分类器

CE 分类交叉熵

BL 添加偏置因子后的分类交叉熵

FL 聚焦损失

P 精确率

R 召回率/查全率

F1 F1 分数

## 2 软件概述

### 2.1 软件功能

（1）输入管理。对输入文本以及用户选择的模型进行记录并显

示。

(2) 模型选择功能。本系统加载了四种基于 Bi-LSTM 的预训练好的因果知识抽取模型供用户选择，它们分别是：Bi-LSTM-CRF，Bi-LSTM-Softmax (CE)，Bi-LSTM-Softmax (BL)，Bi-LSTM-Softmax (FL)。

(3) 序列标注结果生成。系统会加载模型完成一次序列标注过程并将结果显示出来。

(3) 因果知识分析功能。系统会根据序列标注的结果自动分析文本中所存在的因果知识并将结果显示出来。

## 2.2 算法简介

本系统加载了四种基于 Bi-LSTM 的预训练好的因果知识抽取模型供用户选择，它们分别是：Bi-LSTM-CRF，Bi-LSTM-Softmax (CE)，Bi-LSTM-Softmax (BL)，Bi-LSTM-Softmax (FL)。Bi-LSTM 为一种特殊的循环神经网络结构，通过特殊设计的门结构可以有选择地保存上下文信息，其基本结构如图 2-1 所示。

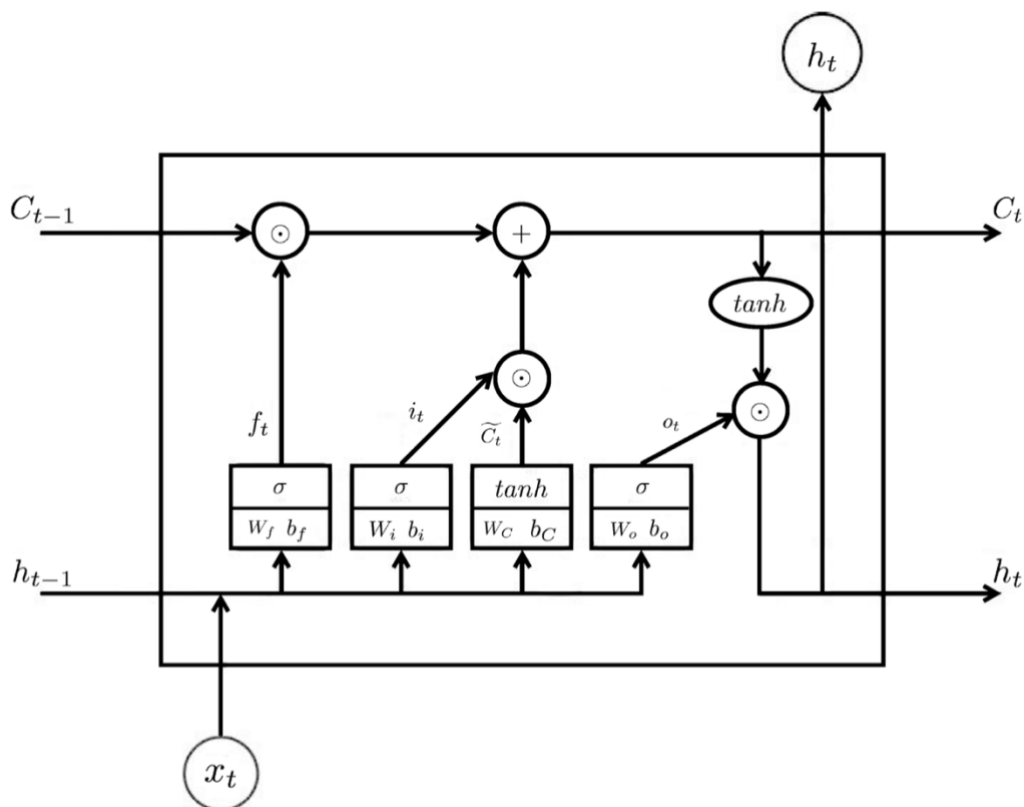


图 2-1 LSTM 单元的结构图

四种模型均为序列标注模型，可解决与下图类似的因果知识抽取任务：

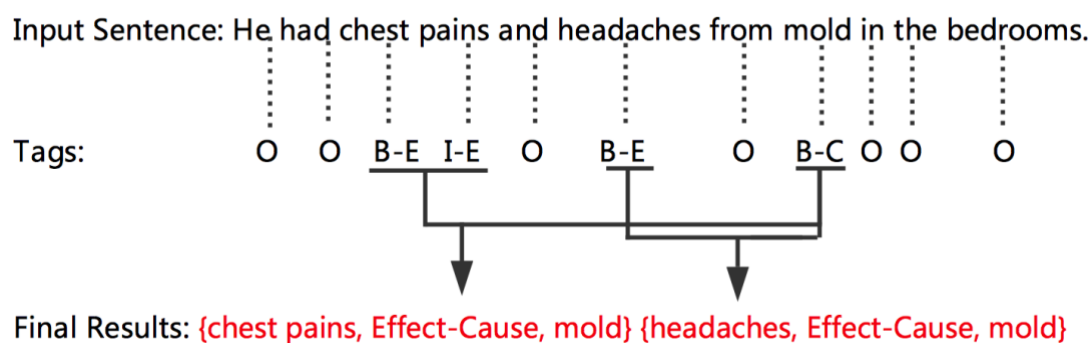


图 2-2 一个典型的因果知识抽取任务

其中，模型 Bi-LSTM-Softmax (FL)的 F1 值和 P 值最高，模型 Bi-LSTM-Softmax (BL)的 R 值最高，Bi-LSTM-CRF 的 P 和 R 值比较均衡，Bi-LSTM-Softmax (CE)的运行速度较快。用户可根据自己的需求选择不同的模型。

下图为模型 Bi-LSTM-Softmax (FL) 的结构图，其他模型的结构与之类似。

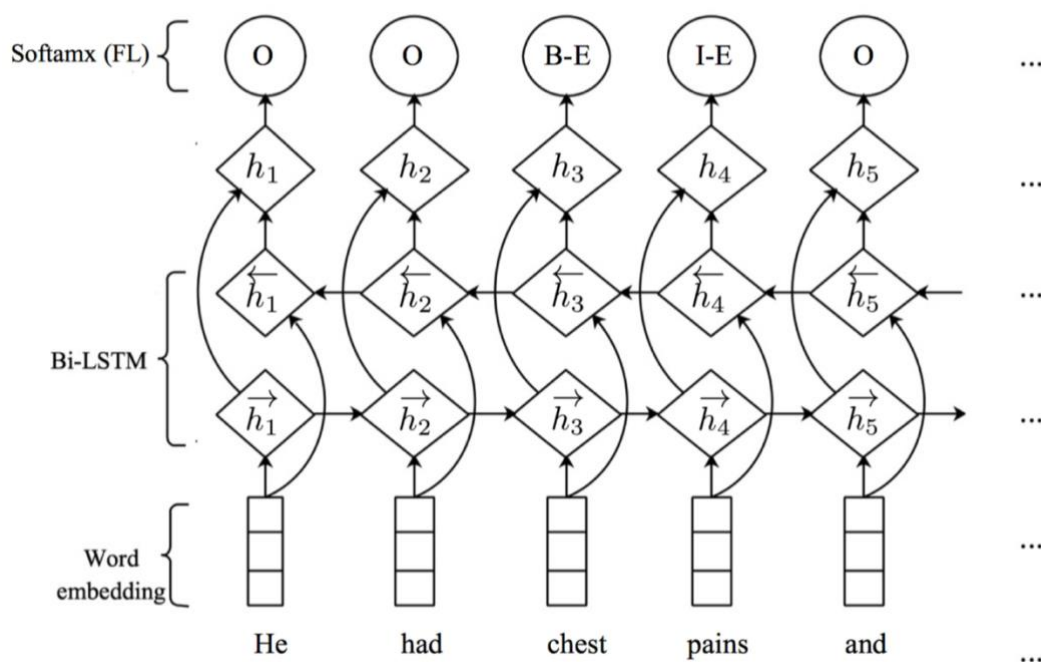


图 2-3 模型 Bi-LSTM-Softmax (FL) 的结构图

## 2.3 软件运行

本系统运行在 PC 及其兼容机上，使用 macOS HighSierra 操作系统，在软件安装后，直接点击相应图标，就可以显示出软件的主菜单，进行需要的软件操作。

## 2.1 系统要求

macOS High Sierra 10.13.3 及以上系统，8 GB 以上内存。

## 3 系统使用

### 3.1 系统启动

在软件安装后，直接点击相应图标，就可以显示出软件的主菜单，进行需要的软件操作。

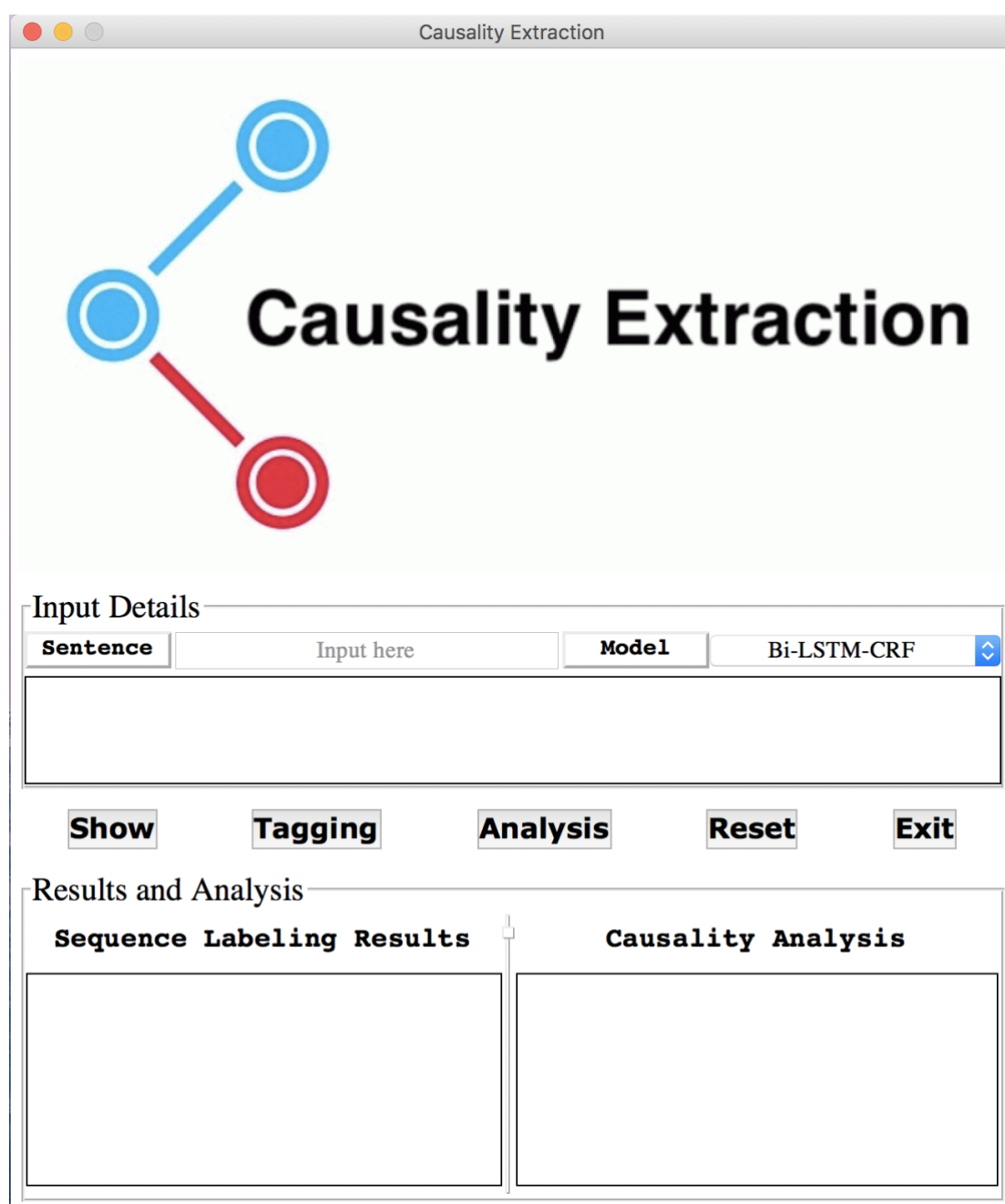
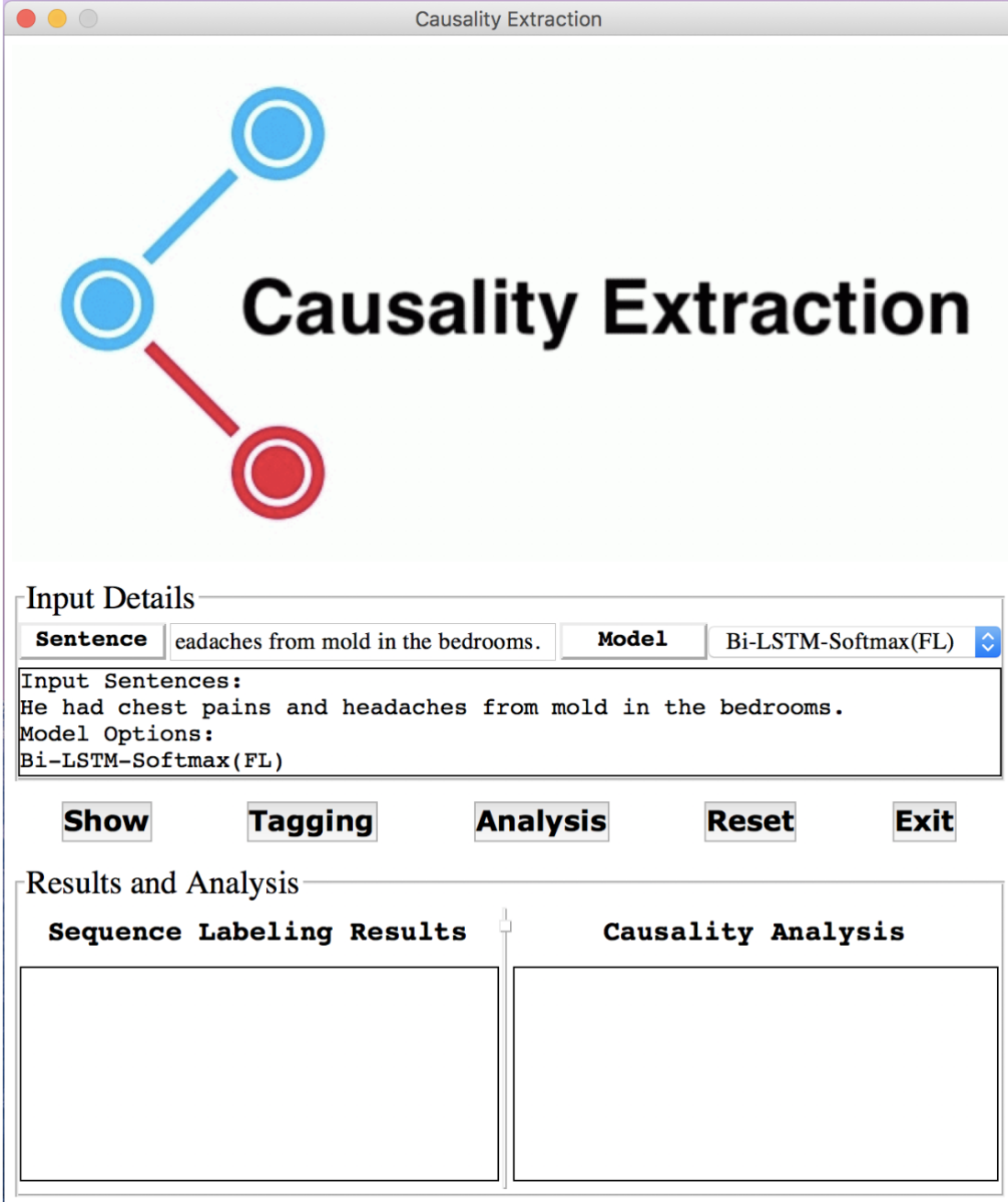


图 3-1 系统启动界面



### 3.2 输入管理

用户需输入有待进行因果知识抽取的句子（目前仅支持英文），并选择模型。



The screenshot shows a web application titled "Causality Extraction". At the top, there is a logo consisting of three blue circles connected by lines, with the text "Causality Extraction" next to it. Below the logo, there is a section titled "Input Details". This section contains a "Sentence" input field with the text "eadaches from mold in the bedrooms." and a "Model" dropdown menu set to "Bi-LSTM-Softmax(FL)". Below these fields, there is a text area displaying "Input Sentences: He had chest pains and headaches from mold in the bedrooms." and "Model Options: Bi-LSTM-Softmax(FL)". At the bottom of the "Input Details" section, there are five buttons: "Show", "Tagging", "Analysis", "Reset", and "Exit". Below the "Input Details" section, there is a section titled "Results and Analysis". This section is divided into two columns: "Sequence Labeling Results" and "Causality Analysis". Both columns have empty rectangular boxes for displaying results.

图 3-2 用户可点击按钮 show 来显示输入细节

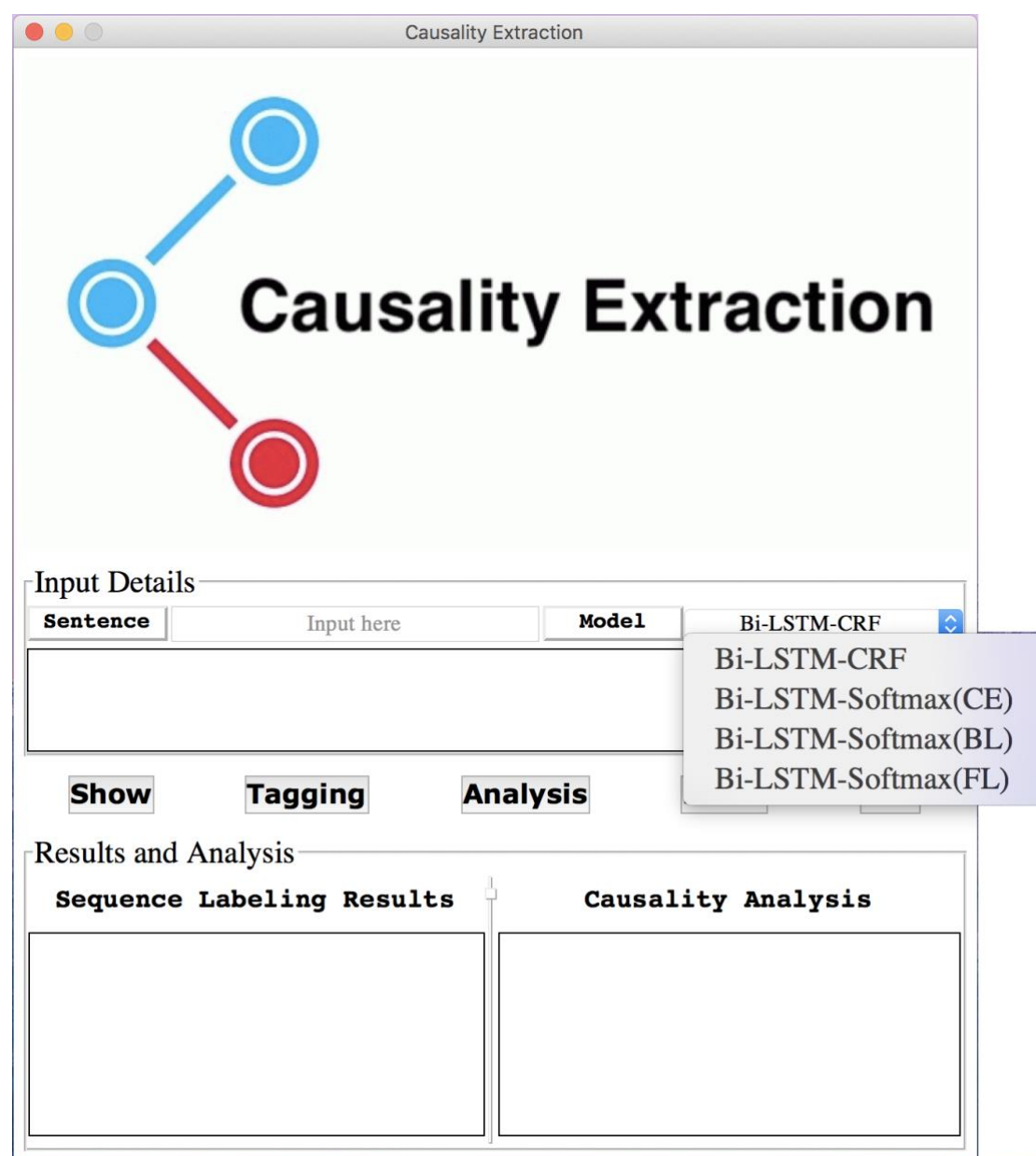


图 3-3 模型选项

### 3.3 序列标注结果生成

用户点击按钮 **Tagging** 后，系统会加载用户所选择的模型对输入语句进行一次序列标注并显示出序列标注结果。

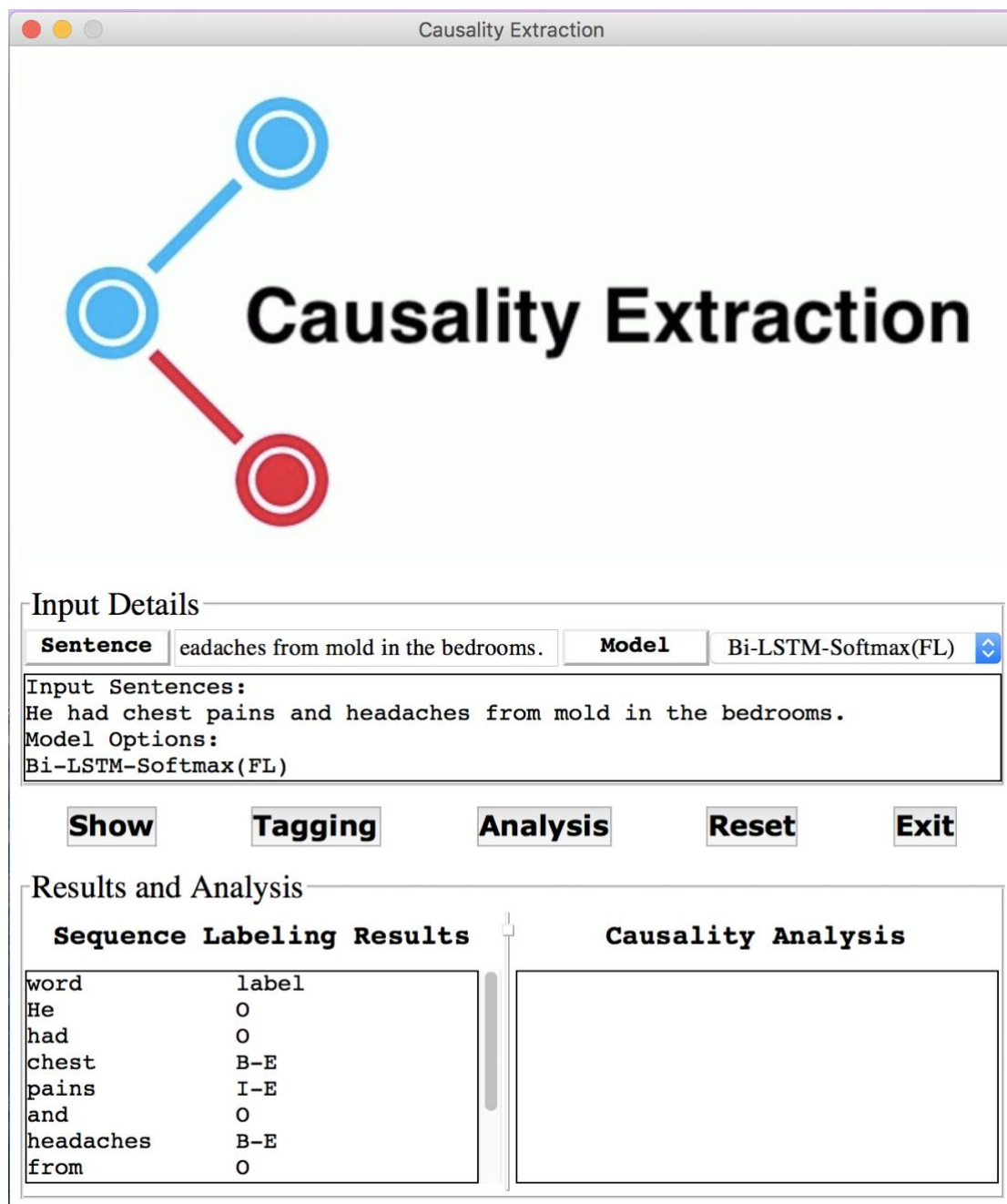


图 3-4 用户可点击按钮 Tagging 来得到序列标注结果

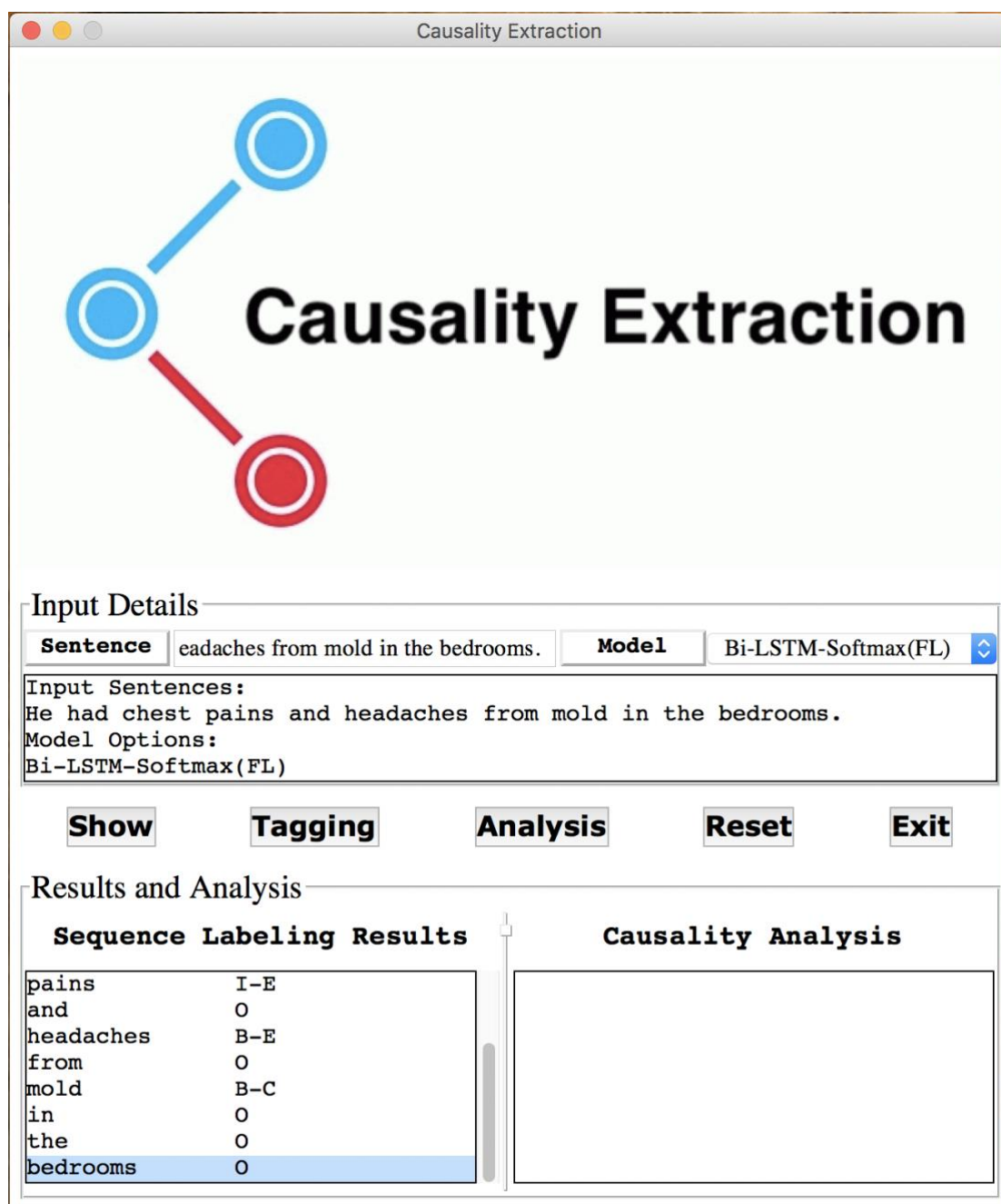


图 3-5 用户可通过滚动条来查看序列标注结果

### 3.4 因果知识分析

用户点击按钮 **Analysis** 后，本系统会根据用户所选模型前向传播生成的序列标注结果，自动分析用户输入自然语言文本中所存在的因果知识，然后将结果显示出来。

最后用户可通过点击按钮 **Reset** 清空序列标注与因果知识分析的

结果，并重置输入英文语句及选择预训练好的模型；用户还可通过点击按钮 **Exit** 来推出本系统。

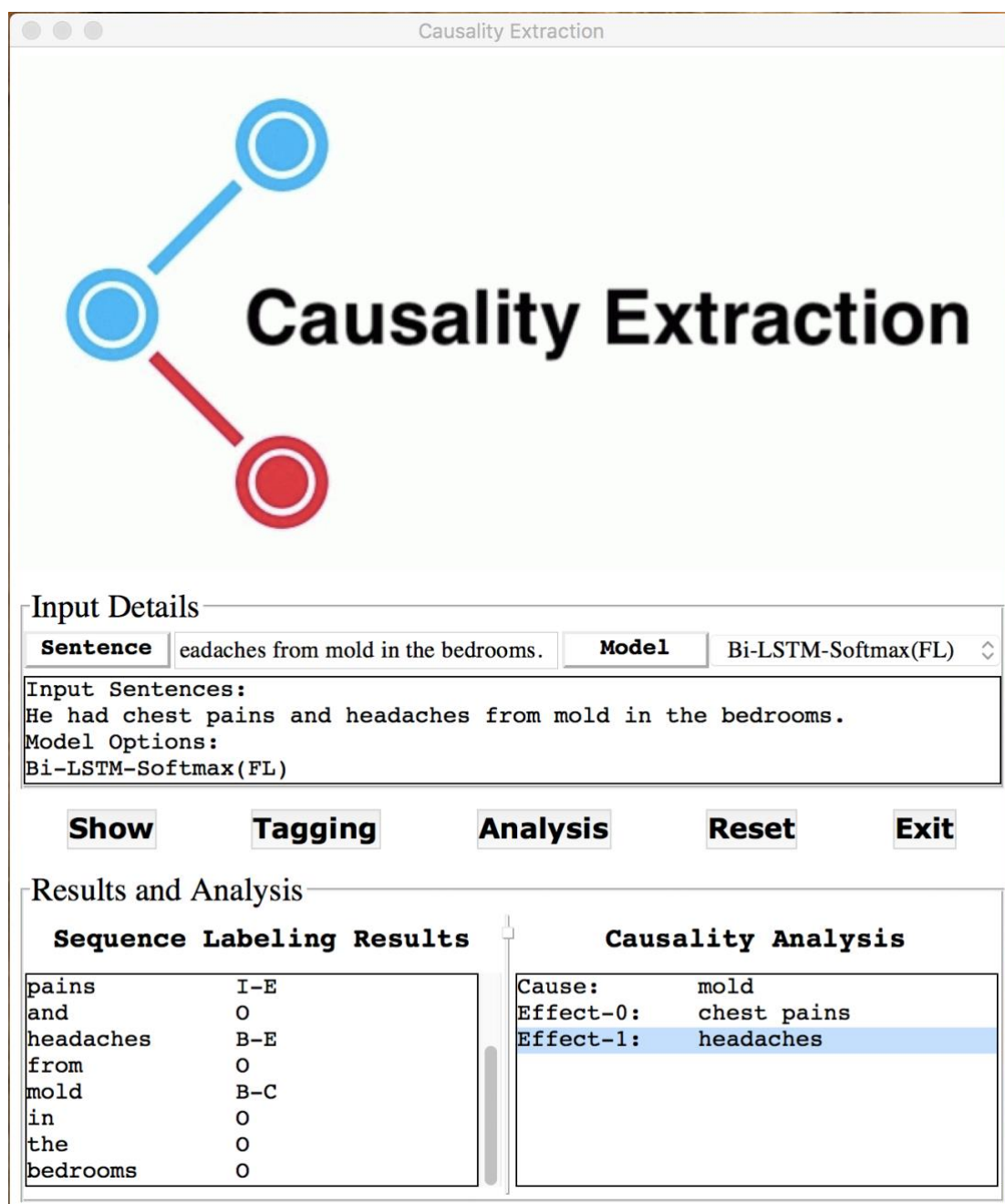



图 3-6 用户可点击按钮 **Tagging** 来得到因果知识分析结果

Causality Extraction



# Causality Extraction

Input Details

Sentence

Input here

Model

Bi-LSTM-CRF

Show

Tagging

Analysis

Reset

Exit

Results and Analysis

Sequence Labeling Results

Causality Analysis

图 3-7 用户可点击按钮 Reset 来清空结果并重置输入