中文时间表达式抽取问题开题报告

张曦
2018年2月

1. 项目背景

T-IR(时间信息抽取)在许多计算机应用中非常关键，如处理"最近好看的电影"、"明天9点提醒我抢票"这样的请求。[1]是对T-IR问题的一篇综述，把T-IR分成了两步：TimEx(时间表达式)的抽取以及Normalization。TimEx抽取是界定句子中，哪些词组成了时间表达式；Normalization则是分析时间表达式具体是指哪个或哪些时候。解决TimEx抽取的方法中，既有基于规则的方法和，也有机器学习的方法；而解决Normalization的方法中，据我所知，都是基于规则的方法。

近年来，关注T-IR问题的评测主要由ACE-2005和TempEval-1/2/3系列。TempEval-2提供了包括中文在内的6种语言的语料库，不过可惜在提交的系统中，没有处理中文语料库的；TempEval-3提供了英语和西班牙语的语料库，且无论规模还是质量，都大大高于TempEval-2。在TempEval-3的T-IR子任务中，基于规则和基于机器学习的方法取得了非常接近的结果[2]。SUTime[3]和HeidelTime[4]是基于规则的系统，其F值都达到了90.3%；ClearTK-TIME[5]和ManTIME[6]是基于机器学习的系统，前者F值达到了90.2%，后者达到了88.1%。

[7]和[8]则是面向中文文本的工作。[7]依然使用的Heidel系统，在[4]的基础上开发了中文规则包，F-score达到89.3%；[8]使用的是机器学习方法，首先对文本进行POS(词性标注)，然后使用CRF(条件随机场)模型进行标注，F-score达到84.7%。

在上面提到的使用了机器学习方法的文章中，都把TimEx标注问题建模为BOI标注问题(Inside–outside–beginning)，而基于神经网络和深度学习的方法在也在标注问题中取得了好的效果。[9]介绍了一种前向神经网络来解决POS标注的方法，虽然不是直接面向TimEx抽取问题，但思路和方法应该可以通用。

本文主要关注时间表达式抽取问题，且仅限于中文文本。

2. 问题描述

TimeML[10]是用于自然语言中事件和时间的标注规范。在TimEx抽取问题中，主要使用TimeML的TIMEX3规范，定义时间表达式为一个或多个时间单元组成的短语，有四种类型：Data、Time、Duration和Set。其中，DATE类型是指一个具体的日历时间，如："公元前456年、上 月十五号、3月"等；TIME指的是一天中的某个时间点或时间段，如："下午3点半，凌晨一点"等；DURATION表示的是一个持续的段时间，如："前几年，两分钟"等；SET描述的是某一类时间的集合，如："每两年，天天"等。TimEx抽取问题不涉及表达式类型分类，仅需要标记出句子中哪些词组成时间表达式。我们使用BOI标记方法对问题建模。中文句子由词组成，而一个TimEx则由1个或多个连续的词组成，因此可以把TimEx的起始词标记为B，后续词标记为I，而非TimEx的词则标记为O。如：

> "今年/B 6月/I ，/O 国家/O 发布/O 声明/O ，/O 要/O 在/O 8月/B 推进/O 这项/O 政策/O 。/O"。

3. 输入数据

有TIMEX标注的中文语料库相对比较少，已知的语料库有TempEval-2和ACE2005语料库，不过ACE2005不能免费获取，因此本文主要使用TempEval-2语料库来训练和评测。该语料库中，训练集包含44篇文档，931个句

子，23180个词，共有763个TimEx；测试集包含8篇文档，195个句子，5313个词，共有131个TimEx。

此外，一些无监督的算法不要求有人工标注，可使用更多的语料库,本文采用1998年人民日报语料库。

4. 解决办法

针对序列标注问题，有很多算法可以应用，如HMM、最大熵算法等。本文将尝试两种方法:

1. 基于CRF的标注方法
2. 基于词向量和神经网络的标注方法

5. 评估指标

采用TempEval-2会议中的评测标准，通过准确率、召回率和F值来评估系统性能。对于预测的结果，针对“词”这一粒度统计其中true positives(tp), true negatives(tn), false positives(fp)和false negatives(fn)的个数。比如，如果人工标注“明天 早上”为TimEx，但预测时，仅认为“明天”为TimEx，那么认为有1个tp和1个fn。

三个指标的计算方法为：

```
precision = tp / (tp + fp)
recall = tp / (tp + fn)
f1-measure = 2 * (precision * recall) / (precision + recall)
```

6. 基准模型

首先基于CRF实现了一个基准模型，该模型仅使用了少量特征模版，如下所示。该CRF标注器的准确率、召回率和F值分别为82.5%、92.8%、74.3%。

> 规则模版：
>
> 1. 当前词
> 2. 前一个词或BOS(begin of sentence)
> 3. 后一个词或EOS(end of sentence
> 4. 当前词的第一个字
> 5. 当前词的最后一个字

此外，项目背景中提到了SUTime、HeidelTime、ClearTK、ManTIME及[8]的系统，他们在TimEx抽取问题的评测结果如下表所示。

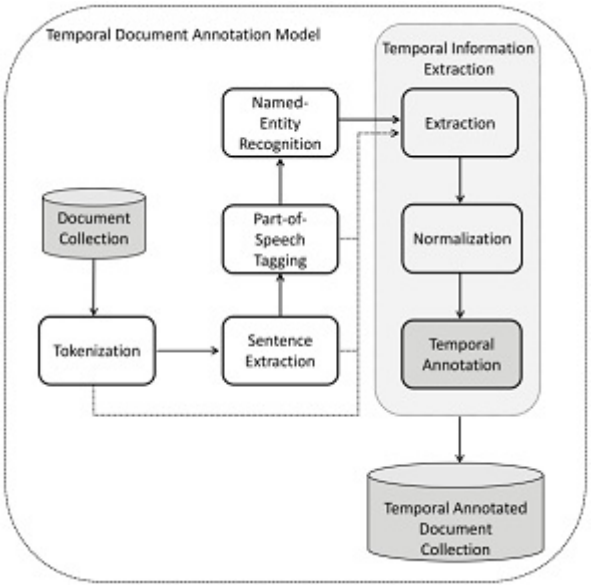
编号	系统	方法	语言	F值	准确率	召回率
1	SUTime	规则	英语	90.32%	89.36%	91.30
2	HeidelTime	规则	英语	90.30%	93.08%	87.67
3	ClearTK	机器学习	英语	90.23%	93.75%	86.96
4	ManTIME	机器学习	英语	89.66%	95.12%	84.78

5	HeidelTime	规则	中文	95.5%	83.8%	89.3%
6	POS-R[8]	机器学习	中文	84.17%	85.16%	83.21%
7	基准模型	机器学习	中文	82.5%	92.8%	74.3%

7. 设计大纲

7.1 CRF标注方法

[1]提到了一个基本的TimEx抽取工作流，如下图所示，分为预处理及抽取两个步骤。



预处理步骤包括：

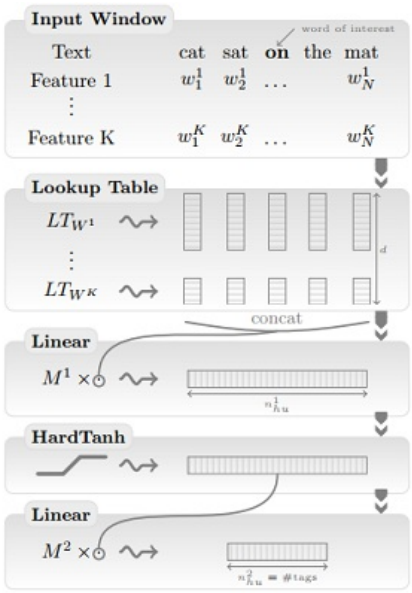
1. 分词
2. POS词性标注
3. NER识别

由于TempEval-2中文语料是已经人工分好词了的，所以可以省略步骤1；NER识别和TimEx抽取在一定程度上，是重合的任务，所以，也可省略步骤3。

在抽取步骤中，首先根据词语特征及POS标注，人工提炼出规则模板，然后输入到CRF模型进行训练。

7.2 基于词向量和神经网络的方法

词向量是对自然文本中词语的一种表示方法，其本身是一个低维的实数向量，以50维和100维比较常见。[9]介绍了一种利用词向量和神经网络进行POS标注的方法，本节主要参考该方法来解决TimEx抽取问题。神经网络结构如下图所示。



该神经网络有三层：线性层、HardTanh非线性层以及输出层。如何通过词语构造神经网络的输入呢？首先把每个词都通过查找表转换为词向量，然后把该词语的其他特征(如POS)进行编码(如独热编码)，结合词向量作为神经网络的输入,词向量则是预先训练好的。这里我们可以通过使用word2vec等工具在人民日报语料库上训练得到词向量。此外还有一点要注意，句子一般是变长的，而神经网络只能处理固定维度的输入，[1]介绍了window方法和sentence方法来处理该问题，这里主要谈window方法。该方法一次仅标注一个词，且假设每个词的标注仅与附近的词相关，那么，我们只需要把该词及周边固定数量K的几个词作为神经网络输入即可，K可作为系统的超参数调节。

8. 附：预计将使用的工具包

1. Stanford POS Tagger
2. pycrfsuite
3. gensim
4. tensorflow

9. 参考文献

[1] Campos R, Jatowt A. Survey of Temporal Information Retrieval and Related Applications[J]. Acm Computing Surveys, 2015, 47(2):15. [2] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 1–9, Atlanta, Georgia, USA,June.
[3] Angel X Chang and Christopher Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In Proceedings of the 8th International Conference on language Resources and Evaluation
[4] Giulio Manfredi, Jannik Strotgen, Julian Zell, and ¨ Michael Gertz. 2014. HeidelTime at EVENTI: Tuning Italian Resources and Addressing TimeML's Empty Tags. In Proceedings of the Forth International Workshop EVALITA, pages 39–43.
[5] Bethard Steven. Cleartktimeml: A minimalist approach to tempeval 2013. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013); Atlanta, Georgia, USA. June; Association for Computational Linguistics;

2013. pp. 10–14.
[6] Michele Filannino, Gavin Brown, and Goran Nenadic. 2013. Mantime: Temporal expression identification and normalization in the tempeval-3 challenge. In Second Joint Conference on Lexical and Computational Semantics.
[7] Hui Li, Jannik Strotgen, Julian Zell, and Michael Gertz. 2014. Chinese Temporal Tagging with HeidelTime. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL '14), pages 133–137. ACL.
[8] 李君婵;汉语文本中的时间、事件及其属性识别[D];山西大学;2013年
[9] 179. R. Collobert J. Weston L. Bottou M. Karlen K. Kavukcuoglu P. Kuksa "Natural Language Processing (almost) from Scratch" J.Machine Learning Research vol. 12 pp. 2493-2537 2011.
[10] Roser Saur´ı, Jessica Littman, et al. TimeML Annotation Guidelines Version 1.2.1 https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_annguide_1.2.1.pdf