北京大学学报(自然科学版) 第 54 卷 第 3 期 2018 年 5 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 3 (May 2018)
doi: 10.13209/j.0479-8023.2017.166

基于多特征的语义角色标注一致性计算方法研究

柯永红[1,†] 朱永福[2] 穗志方[2] 俞士汶[2]

1. 北京师范大学文学院, 北京 100871; 2. 北京大学计算语言学教育部重点实验室, 北京 100871; † E-mail: keyonghong@126.com

摘要 综合语义角色标注语料的格式、标签结构、标签内容和用户可信度等多个特征, 实现一种自动的语义角色标注一致性计算方法。实验证明, 该方法对错误标注的召回率高, 并且速度快, 结果稳定, 可以大幅度地提高语义角色标注一致性检查的效率。

关键词 语料库; 多特征; 一致性计算; 语义角色标注

中图分类号 TP391

A Method for Semantic Roles Labeling Consistency Calculation Based on Multi-features

KE Yonghong[1,†], ZHU Yongfu[2], SUI Zhifang[2], YU Shiwen[2]

1. School of Chinese Language and Literature, Beijing Normal University, Beijing 100871; 2. Institute of Computational Linguistics, Peking University, Beijing 100871; † E-mail: keyonghong@126.com

Abstract The authors state an automatic method for semantic role labeling consistency calculation, based on the features of annotated corpus' format, structure, content and user performances. The expriment shows that the proposed method is fast, stable and has high recall rate, and it can greatly improve the quality and efficiency.

Key words corpus; multi features; consistency calculation; semantic role labelling

语料库是自然语言处理研究和应用的基础资源, 自然语言处理系统的性能和鲁棒性在很大程度上取决于建模过程中是否有足够的高质量标注语料。近年来, 基于深度学习的深度神经网络(deep neutral network, DNN)模型大行其道, DNN 模型更加凸显对大规模、高质量标注语料的强烈需求。

语义角色标注是对句子中的相关体词性成分在谓词表达的事件框架中扮演的语义角色进行标注, 其本质是句子浅层语义分析的一种方法, 在大规模语义知识库的构建、问答系统、机器翻译和信息抽取等领域都有广泛的应用。语义角色标注语料库是自然语言处理研究和应用的基础性资源之一, 高效并可靠的一致性检查是建设大规模、高质量语义角色标注语料库的必要工作。目前, 标注语料一致性检验主要依赖人工。人工检验基本上可以保证标注的准确性, 但是主观性强, 效率低, 代价高昂, 是制约语料标注质量和效率的因素之一。本文尝试基于多个特征来实现语义角色标注一致性的计算, 以期提升语义角色标注一致性检查的速度和质量。

1 相关工作

目前, 针对机器自动计算文本标注语料的一致性的研究成果不多, 仅有的一些研究集中于词义和词性标注方面。虽然针对词层面的一致性自动检查方法[1−5]有许多值得借鉴的地方, 但语义角色标注面对的语料和标注方法更复杂, 直接使用以往的方法不能取得好的效果。

Proposition Bank[6]是目前相对完整、规范的语义角色标注语料库。与英语相比, 中文语义角色标注语料库的研究和建设起步较晚。由于研究目的不

国家重点基础研究发展计划(2014CB340504)资助
收稿日期: 2017−01−20; 修回日期: 2017−11−27; 网络出版日期: 2017−11−28

同，本研究组承担的国家重点基础研究发展计划“融合三元空间的中文语言知识与世界知识获取和组织”项目的语义角色标注有自身的一套标注规范，导致我们不能直接使用现有的语义角色标注语料。我们尝试过基于神经网络的语义角色自动标注方法[7−8]，但受限于训练语料规模，未能达到预期效果。由于一致性检查注重错误标注的召回率，而自动语义角色标注更关注正确率，因此在实际标注过程中，我们使用自动标注方法进行初步标注，在此基础上再由人工标注，最终通过一致性计算来查找可能出错的标注。

为提升标注速度和质量，我们开发了一个协作式标注平台。我们发现，标注过程中的用户行为数据对评价标注一致性有非常重要的影响，如用户在某一类语料上的修改次数越多，或是标注时间越长，说明该类语料的标注难度越大，这类语料的标注结果需要重点关注。用户过往标注的正确率可以反映用户的标注能力。通过用户行为数据分析，可以为标注一致性检查提供非常有价值的参考数据，而这些数据是改进标注系统、推动标注进展的关键。但基于用户行为分析标注一致性，不仅需要好的模型，更需要大量而详尽的用户数据。鉴于目前的研究条件尚不充分，我们在本文方法中加入初步的用户标注可信度计算。

语义角色标注的一致性计算既有重要的研究价值，又有广泛的工程应用前景。但是，目前对语义角色标注一致性计算的研究不足。因此，本文提出基于多特征的语义角色标注一致性计算方法。

2 基于多特征的语义角色标注一致性自动检验方法

语义角色标注一致性计算的目标是根据标注规范、标注文本特征和用户行为数据，为标注结果计算一致性打分，减轻人工检查的工作量，提升标注质量。图 1 描述该模型的流程。

从图 1 可以看出，模型的执行过程包括以下几步：1) 对输入的标注语料进行格式检查；2) 对输入的标注语料进行结构检查；3) 对输入的标注语料进行内容检查；4) 根据格式、结构和内容检查结果，生成错误 id 字符串；5) 根据以往的修改记录，计算该用户的可信度；6) 根据用户可信度和错误 id 字符串，生成该条标注的一致性得分，进行打印输出。

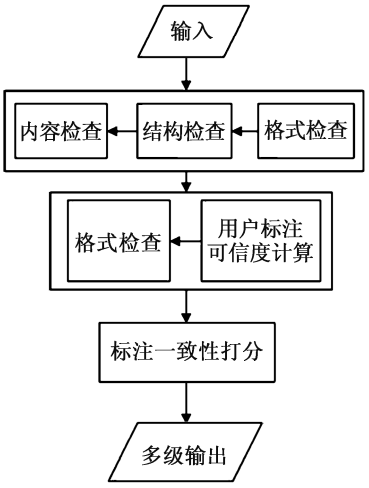

图 1 基于多特征的语义角色标注一致性计算模型
Fig. 1 Model of SRL calculation based on multi-feature

2.1 输入语料

模型的输入为经过语义角色标注的文本及相关附加信息，其格式如表 1 所示。

2.2 格式检查

格式检查主要依据标注规范，检查标注文本的基本格式，包括标注文本是否修改了原始语料，标签格式是否正确(符号配对和多余空格)，标签是否符合操作规范的定义，等等。格式正确是最基本的要求，格式错误会导致标签内容提取错误。因此，一旦检查到格式错误，算法会停止检查，直接输出错误信息。如果格式检查正确，则返回的错误 id 为 0。格式检查能够发现的 4 种错误如表 2 所示。

2.3 标签检查

标签检查主要包括数量检查、标签互斥关系检查和标签依存关系检查 3 个方面，可以检出的错误如表 3 所示。

2.3.1 数量检查

语义角色标注是对句子中的相关体词性成分在谓词表达的事件框架中所扮演的语义角色进行标注，因此每条标注有且仅有一个谓词。如果标注文本未出现谓词标记或出现多个谓词标记，则判定为错误的标注。表 4 列出部分示例。

2.3.2 互斥性关系检查

互斥关系检查主要包括以下两个方面。

1) 一条标注中同一标签至多出现一次，如果一个谓词有多个论元的论旨角色相同，则应当采用[%+ %]或[%& &]标签来辅助标注。[%+ %]用来标记部分论元成分，[%& &]用来标记同指论元

表 1　输入语料格式
Table 1　Format of input corpus

字段名	内容	举例
标注 id	唯一标识该条标注结果的字符串	402880dd513330bb01514990901101be
原始语料文本	分词后的文本	作为 一 名 警察 ， 我 只能 对 薛 永清 夫妇 的 离世 表达 沉痛 的 哀悼 。
标注文本	对分词后的语料添加相应的语义角色标签后的文本	作为 一 名 警察 ， 我 只能 [%对象 对 薛 永清 夫妇 的 离世 %] [# 表达 #] [%受事 沉痛 的 哀悼 %] 。
修改标记	状态: 存疑/修改/删除/空白; 次数: 该标注被修改的次数	2
原标注 id	基于原标注结果 id 对应的标注文本进行修改	
标注用户 id	唯一标识该标注用户的字符串	xiaoqiang@pku.edu.cn
标注时间	用户标注该文本的时间	11/27/2015 23:29:49
原始语料 id	原始语料的 id	402880dd513330bb01514990900201bb

表 2　格式检查的错误类型
Table 2　Types of format checking

序号	错误信息
1	原句文字被修改
2	多余的空格
3	标签符号不匹配
4	错误的标签名称

表 3　标签检查的错误类型
Table 3　Types of label checking

序号	错误信息
1	谓词个数错误
2	错误的标签组合
3	标签重复
4	缺少$(标签名)标签
5	可能需要$(标签名)标签

表 4　数量检查示例
Table 4　Example of predicatenumber checking

标注示例	检验结果
开 [%内容 调查会 %] 人 多 好, 还是 人 少 好?	错误
开 调查会 [%当事 人 %] [# 多 #] 好, 还是 人 少 好?	正确
[# 开 #] 调查会 [%当事 人 %] [# 多 #] 好, 还是 人 少 好?	错误

成分。下面两种标注是正确的:

① [%施事 他] 将 陪同 [%+施事 美国人] [# 访问 #];

② [%&与事 主任 %] 要 [%施事 我 %] 随时 [%内容 把 有关 情况 %] [# 通知 #] [%与事他 %]。

2) 部分不同标签之间彼此不能共存，如具有从属关系的标签“时间”和“时段”不能同时出现。

2.3.3　依存关系检查

大部分标签可以独立使用，少数标签不能独立使用，必须依赖于独立的标签。例如“同事”的角色全称是“共同施事”，标注那些需要两个或两个以上的施事共同完成的谓词，它的存在必须要在“施事”角色已经存在的前提下:

① [%施事 代表们%] [%同事 和厂长%] 进行了[# 谈判 #];

② 其实 [%施事 他%] 也是看中[%同事 和中国大陆%] [# 做生意 #] 的机会。

2.4　内容检查

内容检查就是检查标签角色与标签标记的文本内容语义是否一致，是语义角色标注一致性计算的重点内容。我们通过计算标注文本和标注范例之间的余弦相似度，基于 tf-id 来计算语义角色一致性，这种方法检验出的结果依赖于训练数据的覆盖率和准确率。内容检查的流程如图 2 所示。

余弦相似度指利用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异，余弦值越接近 1，表明夹角越接近 0°，也就是两个向量越相似，其计算公式为

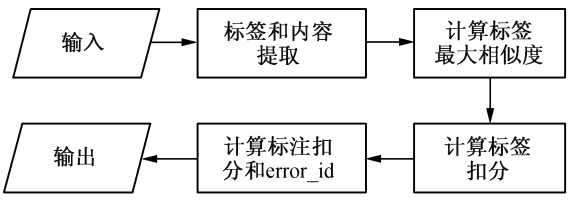

图 2　标注语料内容检查流程
Fig. 2　Flow chat of content checking

$$\cos\theta=\frac{\sum_1^n A_i\times B_i}{\sqrt{\sum_1^n A_i^2}\times\sqrt{\sum_1^n B_i^2}}。\quad(1)$$

对于输入数据中的语义角色标签 X, 可以将其标注的语义内容视为由一个个切分后的最小粒度 terms 的 tf-idf 相似度组成的向量 ***A***。提取训练数据中该语义角色标签 X 标注下的所有内容, 可以得到待比较向量 ***B*** 的集合, 获取向量 ***A*** 与向量 ***B*** 的最大相似度, 即可作为标签 X 标注的语义内容的检查结果。

为了计算每条 terms 的 tf-idf 值, 需要获取这些 terms 的 df 词典和 tf 值。terms 的 df 值为训练数据中该 terms 出现的次数, tf 值是该条标注中该 terms 出现的次数。计算 tf-idf 时使用加 1 平滑, 这样, 一条 terms 的 tf-idf 值就是

$$\mathrm{tf\_idf[terms]}=(\mathrm{tf[terms]}+1)/(\mathrm{df[terms]}+1)。\quad(2)$$

得到每条标签的相似度后, 还需进行一致性的综合计算, 以此作为该条标注在这一步检查中的最终扣分。对于一条标注 ***A***, 假设其标签包括 a_1, a_2, …, a_n, 将单个标签 a_i 标注的语义内容的最大相似度定义为 max_similarity [a_i], a_i 的扣分定义为 score[a_i], 它与 max_similarity[a_i]之间有如下的函数关系:

$$\mathrm{score}[a_i]=\begin{cases}0, & 当\mathrm{max\_similarity}[a_i]>0.95,\\ (1-\mathrm{max\_similarity}[a_i])\times 5, & 当0.8<\mathrm{max\_similarity}[a_i]\leqslant 0.95,\\ (1-\mathrm{max\_similarity}[a_i])\times 10, & 当0.5<\mathrm{max\_similarity}[a_i]\leqslant 0.8,\\ (1-\mathrm{max\_similarity}[a_i])\times 40, & 当0.1<\mathrm{max\_similarity}[a_i]\leqslant 0.5,\\ 60, & 当\mathrm{max\_similarity}[a_i]\leqslant 0.1)。\end{cases}\quad(3)$$

A 的最终扣分 score[***A***]的计算公式为

$$\mathrm{score}[\boldsymbol{A}]=\begin{cases}\left(\sum_{i=1}^n \mathrm{score}[a_i]\right)/n, & \\ \quad \forall a_i,\ \mathrm{max\_similarity}[a_i]>0.1, & \\ 60, \text{ 其他。} & \end{cases}\quad(4)$$

采用这样的非线性策略计算 score[a_i], 是想突出那些存在很大错误嫌疑的标签, 因为一旦出现最大相似度小于 0.1 的标签, 就极有可能是错误的标注。***A*** 的最终扣分 score[***A***]的计算也是出于同样的“木桶原理”, 如果出现相似度极低的标签, 那么这条标注会归类到错误的结果中。

2.5 用户可信度计算

输入中的用户 id 字段标记了该条标注出自哪个用户, 据此可以分析这个用户在某批数据标注中的整体表现, 获取该用户的可信度得分。该可信度得分可用于该用户的所有标注。

对特征的检查结果, 在算法中用 error_id 进行记录。一条标注的错误 id 字符串是后续用户可信度和标注最终得分计算的重要参数, 在算法中是由[0, 5]组成的字符串(形如“0453”), 由每一步对特征检查得到的错误 id 拼接而成, 具体描述如表 5 所示。

如果该用户在某次标注中表现很差, 说明该用户在标注这批数据时的状态可能不大好, 或者该用户标注能力有限。出于召回错误标注的考虑, 需重新检查该用户标注的数据。因为不能排除用户能力提高和状态调整的情况, 所以单一批次标注数据的检查结果不影响该用户的下批数据标注。

根据标注特征检查结果(即返回的错误 id)进行加权处理, 结果如表 6 所示。用户最后的可信度是该标注错误 id 字符串的加权平均值, 其计算公式如下:

$$\mathrm{user\_trustworthiness}=\left(\sum_{\mathrm{error_{id}}}\mathrm{weight}[\mathrm{error_{id}}]\right)/\mathrm{sizeof(error\_id[\ ])}。\quad(5)$$

2.6 标注一致性分级

我们用得分值表示一致性检查结果。首先, 根

表 5 error_id 的含义

Table 5 Meaning of error_id

error_id	含义
0	没有错误
1	格式上的错误, 一旦发生算法立刻终止, 对用户可信度无影响
2	格式上的错误, 一旦发生算法立刻终止, 对用户可信度影响很大
3	标签结构或者标签语义角色一致性上的重大错误, 完全不符合语义角色标注规范, 一旦发生算法立刻终止
4	标签结构或者标签语义角色一致性上的非重大错误, 会影响标注得分和用户可信度, 发生后算法仍会继续检查其他特征
5	标签结构或者标签语义角色一致性上的疑似错误, 会轻微影响标注得分和用户可信度, 发生后算法仍会继续检查其他特征

据错误 id 字符串计算初始扣分，结果如表 7 所示；然后，将初始扣分加上内容检查环节扣分，得到最终扣分；接着，在基础分 100 中(如果该条标注之后被别人修改了，则基础分为 60)减去最终扣分；最后，乘以用户可信度，得到最终得分。最终得分是一个不小于 0 的浮点数，计算公式为

最终得分= (基础分−错误 id 扣分−内容检查扣分) × 用户可信度。

将最终得分映射到一致性分级，结果如表 8 所示。为了保证召回率，将“可能错误”的区间范围设置最大。

2.7　错误的多级输出

输出字段包括标注结果 id、错误 id 字符串、一致性分级和多级拼接的错误信息。其中，多级拼接的错误信息字段是根据错误 id 字符串对应得到的。理论上，不为 0 的错误 id 都会对应一条错误信息，反映该标注结果的出错原因或者可能存在的问题，对后续的人工修改或者人工二次检验具有参考价值。将表 1 中的标注文本作为样例，输出的结果如表 9 所示。

从表 9 可以看出：1) 输出中错误 id 字符串为“00553”(表示在特征检查过程中出现 error_id 为 3 和 5 的错误)，正确性分级为第 5 级；2) 多级拼接错误信息字段提示该标注可能存在的问题，包括“对象”标签可能缺少对应的当事或者施事标签、对象标签与受事标签与训练数据的相似度太小等问题。

表 6　不同 error_id 对应的权重值

Table 6　Corresponding between error_id and weight

error_id	权重	error_id	权重
0	1	3	0.25
1	1	4	0.5
2	0	5	0.9

表 7　error_id 对应的初始扣分

Table 7　Corresponding between error_id and initial score

error_id	score	error_id	score
0	0	3	80
1	100	4	40
2	100	5	5

表 8　一致性分级

Table 8　Grade of consistency calculation

一致性分级	含义	最终得分
1	完全正确	(95, 100]
2	正确	(80, 95]
3	不确定	(60, 80]
4	可能错误	(20, 60]
5	错误	[0, 20]

3　实验结果与分析

3.1　实验语料

目前，国内外都没有语义角色标注的一致性计算公共测试集，我们采用国家重点基础研究发展计划项目的部分语料。课题的语义角色标注由北京大学负责制定标准，用于实验的部分语料情况如表 10 所示。标注例句集 1 经过多次校对，标注质量较高。我们抽取其中 8000 句作为内容检查的训练语料，其余部分作为测试语料。

3.2　实验结果评价标准

我们采用召回率和准确率两个指标来衡量实验结果。

1) 首先抽取 ICL 平衡例句集中 8000 句作为内容检查的训练语料，然后对实验集分别进行一致性计算，保存计算结果。设实验中一致性得分小于 4 (即认为不确定、可能错误、错误都需要人工检查)的句子数为 $\mathrm{SelectedSen}_i$，由人工检验被修改的句子数为 M_i，$\mathrm{SelectedSen}_i$ 中与人工检验结果一致的句子数为 $\mathrm{CorrectSen}_i$。

2) 一致性计算的目标在于筛选存在错误的标注，并提示可能存在的错误之处。因此，召回率是

表 9　输出结果样例

Table 9　Example of results

输出字段	内容
标注结果 id	402880dd513330bb01514990901101be
错误 id 字符串	00553
一致性分级	5
多级拼接的错误信息	可能需要施事或者当事标签|谓词: 1.0|受事: 3.773923848E-7|对象: 0.0043632069

表 10　实验语料集

Table 10　Corpus of the experiment

内容	总句数	总字数	备注
ICL 平衡例句集	10629	124074	人民日报 2000 年 1 月语料。经过人工多次详细检验，一致性有保证
哈工大语料	3231	72964	经过一次人工检验
微博语料	3221	46259	经过一次人工检验
人民日报 1998 年 1 月语料	74000	2800000	多领域、多体裁 150 万字语料标注，部分经过人工检验

最重要的指标，计算公式为

$$R_i = \frac{\text{CorrectSen}_i}{M_i} \times 100\% \text{。} \tag{6}$$

3) 将标注语料中的最终修改与 error_id 提示的错误进行对比，若 error_id 提示的错误类型与最终修改一致，记为 CorrectSen_i，准确率的计算公式为

$$P_i = \frac{\text{CorrectSen}_i}{\text{SelectedSen}_i} \times 100\% \text{。} \tag{7}$$

3.3 实验结果与分析

准确率的计算需要人工逐条比对语料的修改部分，我们在每个实验语料集中分别挑选 200 句，进行准确率的计算。此外，只有 ICL 平衡语料集部分有用户修改记录，进行过用户可信度打分，其他实验语料集没有用户修改数据，计算过程中默认用户可信度为 1。实验结果如表 11 所示，可以得到以下结论。

1) 由于一致性计算的主要作用是挑出可能的错误标注，因此召回率是最重要的指标。我们的模型在上述 4 个实验集语料中取得较高的召回率，具备不错的“挑错”能力。

2) 一般来说，标注句长越短，句子语义角色标注难度越低；句子越长，标签间的组合关系越复杂，错误输出部分准确率越低。ICL 平衡例句集的平均句子长度最短，且最初由人工在文本上直接标注，容易检出的格式问题较多，所以准确率最高；微博语料次之；人民日报 1998 年 1 月语料平均句子长度最大，得到的准确率最低。通过对比程序和实验结果，我们发现格式检查和标签检查主要依赖标注规范，此部分程序使用了很多规则，检查结果相对准确。内容检查部分对训练语料的规模和代表性要求较高，由于训练语料较稀疏，导致区分 error_id 值 3, 4 和 5 的正确率较低。

3) 此外，由于课题的标注规范处于完善的过程中，对正确率也有影响，如标签“对象”和“内容”，在标注例句“[%施事 爱 夸张 事实 的 孩子 %] 往往 [# 喜欢 #] [%对象 喜剧 %]”和“[%施事 爱 夸张 事实 的 孩子 %] 往往 [# 喜欢 #] [%内容 喜剧 %]”中，很难区分哪个正确。

表 11 实验结果

Table 11 Result of the experiment

语料名称	召回率/%	准确率/%	平均句子长度
ICL 平衡例句集	87.71	61.20	11.67
人民日报 1998 年 1 月语料	89.25	46.25	37.84
哈工大语料	91.64	53.33	22.58
微博语料	84.97	49.28	14.36

4 结论

本文提出基于多特征的语义角色标注一致性计算方法。该方法先从格式、标签结构、标签内容 3 个方面对标注语料进行检查，并输出错误信息，然后根据用户以往的标注情况，计算用户的标注可信度，最终输出一个一致性得分。本文挑选 4 个标注语料集合进行实验，结果显示，本文方法召回率较高，对筛选错误标注有很好的参考价值，大大减轻了人工检查的工作量。但是，所提算法中对错误的提示部分还需要进一步改进，如采用比 tf-idf 更好的相似度计算方法。下一步，我们将进一步丰富训练语料，改进错误分级算法，为修改错误提供更详细的信息。另外，语义角色标注一致性计算方法如何与自动标注程序更有机地相结合，以期提升自动标注的质量，也是我们今后重点研究的方向。

参考文献

[1] 张虎. 汉语语料库词性标注一致性检查及自动校对方法研究[D]. 太原: 山西大学, 2005

[2] 钱揖丽, 郑家恒. 汉语语料词性标注自动校对方法的研究. 中文信息学报, 2004, 18(2): 30−35

[3] 乔剑敏, 张仰森. 词义标注一致性检验系统的设计与实现. 中文信息学报, 2010, 24(4): 44−51

[4] 任丽君. 汉语虚词用法标注一致性检测研究[D]. 郑州: 郑州大学, 2013

[5] 刘江, 郑家恒, 张虎. 中文文本语料库分词一致性检验技术的初探. 计算机应用研究, 2005, 8(22): 52−54

[6] Xue Nianwen, Palmer M. Annotating the propositions in the Penn Chinese Treebank // Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. Stroudsburg, PA, 2003: 47−54

[7] Sun Weiwei, Sui Zhifang, Wang Meng, et al. Chinese semantic role labeling with shallow parsing // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, 2009: 1475−1483

[8] 王臻, 常宝宝, 穗志方. 基于分层输出网络的汉语语义角色标注. 中文信息学报, 2014, 28(6): 56−61