

文本蕴涵的类型和识别机制

——从认知假设到计算模型

北京大学中文系

袁毓林

2007年11月6日

1. 信息检索碰到的难题

- a. Who is Ariel Sharon?
- b. Israel's Prime Minister, Ariel Sharon, visited Prague. — T
- c. Ariel Sharon is Israel's Prime Minister. — H
- 对于给定的问句a来说，文本b蕴涵了句子性假设c；因此，b可以作为a的答句。

2. 文本蕴涵的一般定义

- 语义表达形式的多样性：在自然语言中，相同的意义可以用不同的形式来表达，或者从不同的文本中推断出来。例如：
 - a. Guerrillas killed a peasant in the city of Flores. — T
 - b. Guerrillas killed a civilian. — H
- 可以把这种不同文本之间的同义和蕴涵关系系统称为文本蕴涵（textual entailment）。

3. 文本蕴涵的形式定义

- 文本蕴涵可以定义为：一个连贯的文本（text） T 和一个被看作是假设（hypothesis） H 之间的一种关系。
- 如果 H 的意义（置于 T 的语境中解释时）可以从 T 的意义中推断出来；那么我们说 T 蕴涵（entail） H （即 H 是 T 的推断），记作 $T \Rightarrow H$ 。

4. 为语义表达的多样性建模

- 需要为语言中的这种表达多样性建立模型，以便应用系统能够识别一个特定的目标意义可以从不同的文本变异形式中推断出来。
- 这种模型可以在浅层的语义平面上建立，比如，基于不同的文本表达之间的蕴涵关系这种观念，通过直接对有关的词汇-句法单位进行操作，来发现概率性的语义推理规则。
- 这样，通过为不同的语言表达形式之间的蕴涵关系建立通用的模型，指定一个语言表达形式的意义可以从另一个语言表达形式推断出来的条件；来发展一种识别语言表达多样性的技术路线，从而为多种语言处理应用服务。

5. 文本蕴涵的类型

- 识别文本蕴涵需要综合运用语言知识、世界知识和逻辑推理等多方面的知识；
- 可以根据这些知识的层面分出文本蕴涵的不同类型，揭示其背后的推理机制，并确定识别不同类型的文本蕴涵的难度等级。
- 词汇和词汇-句法两种层面的蕴涵；
- 语义和逻辑层面的蕴涵。

6. 词汇层面的蕴涵的推理机制

- 在词汇层面上，可以假定：文本 T 和假设 H 都是由一组词语表达的，暂时忽略虚词。
- 于是，如果 H 中的每一个词语 h 能够跟 T 中相应的主蕴涵（entailing）词语 t 相匹配，那么 T 和 H 之间具有蕴涵关系。具体地说，如果 h 和 t 共有相同的词目（lemma）和词类，或者通过一系列词汇转换， h 可以跟 t 相匹配；那么，可以认为 t 蕴涵 h 。
- 这样，词汇之间的语义和转换关系，成为人们了解词汇层面上蕴涵关系的推理机制的钥匙。

7. 建立蕴涵型式以便推理

- 发现一对（组）看上去是描述大致相同的事实的相匹配的文本片段，并找出共同的词汇项目作为一组“支撑点”（anchors）。那些跟已知的支撑点共有相同的关系的相应的成分，被习得为互释型式（paraphrase patterns）。例如：
 - a. Yahoo bought Overture.
 - b. Yahoo own Overture.
- 可以从中演绎下面这种蕴涵型式：
$$X \leftarrow \text{subj buy obj} \rightarrow Y \Rightarrow X \leftarrow \text{subj own obj} \rightarrow Y$$
并把Yahoo和Overture作为支撑点。

8. 词汇-句法层面的蕴涵的推理机制

- 在词汇-句法层面上，可以假定：文本 T 和假设 H 都是由一组可通过依存关系分析而得到的句法依存关系来表达的。
- 于是，如果 H 中的关系可以被 T 中的关系所覆盖，那么定义 T 和 H 之间具有蕴涵关系。在有的情况下，如果构成 H 所有关系逐词地出现在 T 中，那么可以认为 T 和 H 之间在词汇-句法层面上具有蕴涵关系。在其他情况下，这种覆盖可以这样来获得：通过对 T 中的关系实施一系列转换，使之能够产生出 H 中的所有关系。

9. 基于依存树匹配关系的识别机制

- 工作原理：寻找文本和假设的依存关系分析树之间的匹配关系。
- “文本-假设”对之间是否具有蕴涵关系，是根据它们之间的相似性来决定的。而这种相似性被定义为假设中能够跟文本匹配上的节点的比例。
- 经过试验，发现这种相似性的阈值是50%。也就是说，当假设中能够跟文本匹配上的节点的比例达到或超过50%时，可以说它们之间具有蕴涵关系；当这种比例小于50%时，可以说它们之间不具有蕴涵关系。

10. 利用谓词-论元结构的识别机制

- 工作原理：用简单的一般性的启发式和知识贫乏的方法来识别同义互释，用NP同指互参、NP语块切分、RASP和Link两个分析器来给“文本-假设”对中的每一个句子产生（PAS）。
- 然后，用WordNet词汇链和一些专门的启发式规则来建立这些PAS中相应成分的语义相似性；
- 最后，为这些相应的PAS的结构相似性和相应词汇成分的相似性设定阈值，用以判断“文本-假设”对之间的蕴涵关系是否成立。

谢谢大家！