文本到文本的生成

文本到文本的生成是指对给定的文本进行变换和处理从而得到新文本的技术。具体可以分为：文本摘要、句子压缩、句子融合、文本复述。

本文主要介绍下句子压缩和融合、以及文本复述。

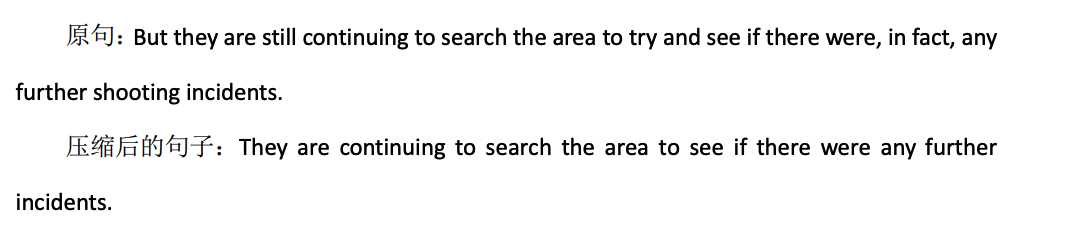
参考地址：

<https://wanxiaojun.github.io/TextGenerationSurvey.pdf>

1. 文本压缩和句子融合

文本压缩和句子融合技术一般用于文本摘要系统中，用于生成信息更加紧凑的摘要，获得更好的摘要效果。

句子压缩技术基于一个长句子生成一个短句子，要求改短句保留长句中的重要信息，也就是重要信息基本不损失，同时要求改短句是通顺的。下面给出一个句子压缩的例子：

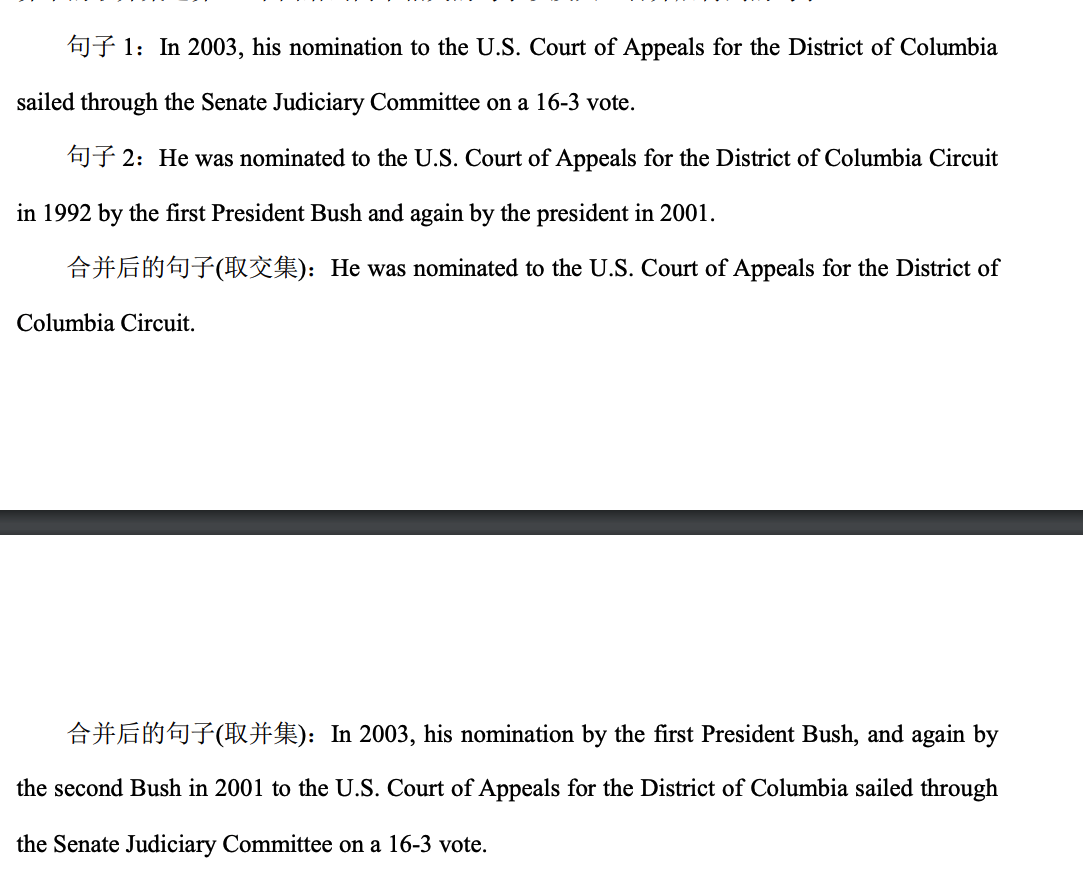


学术界尝试了多种方法实现句子压缩，包括从句子中删除词、对句子中的词语进行替换、重排序或者插入。其中，从句子中直接删除词的做法因其复杂度较低而成为主流方法。

研究人员提出了多种方法用于实现基于词语删除的句子压缩，包括噪声信道模型【1】、结构化辨别模型【2】、树到树的转换【3】、整数线性规划【4】等。但总体效果而言，压缩效果并不明显。

句子融合技术则是合并两个或多个包含重叠内容的相关句子得到一个句子。融合分为两种：一类是只保留多个句子中的共同信息，过滤掉无关的细节信息，另一类是过滤掉多个句子中的重复内容。

下面是一个句子融合的例子：

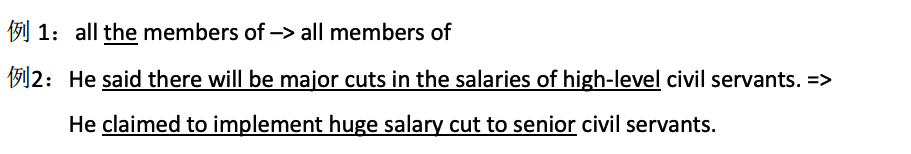


1. 文本复述

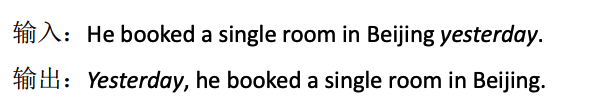
文本复述生成文本通过对给定文本进行改写，生成新的复述文本，一般要求输出文本与输入文本在表达上有所不同，但所表达的意思基本一样。

文本复述生成技术应用很广泛，例如，在机器翻译系统中可以利用文本复述技术对复杂输入文本进行简化从而方便翻译；在信息检索系统中可以利用文本复述技术对用户查询进行改写；在儿童教学中可以利用文本复述技术将难以理解的文本简化为儿童容易理解的文本。

根据实际的需求，通过复述生成技术得到的输出文本与原文本相比，可以是一两个词发生了改变（如例1），也可以是整段文本面目全非（如例2）：



简单的文本复述可以通过同义词替换来实现，也可以通过人工或自动构建复述规则来实现【5】,例如根据变换状语位置的一条规则，可以获得下面句子的简单复述句子：



为了实现复杂的文本复述，研究人员提出了基于自然语言生成的方法【6】、基于机器翻译的方法【7】、基于支点(Pivot)的方法【8】【9】等。

* 基于自然语言生成的方法：模拟人类的思维方式，首先对输入句子进行语义理解，获得该句子的语义表示，然后基于得到的语义表示生成新的句子。
* 基于机器翻译的方法：将文本复述生成问题看作是单语言机器翻译问题，从而利用现有机器翻译模型（例如噪声信道模型）来为给定文本生成复述文本。
* 基于支点的方法：把当前语言中的输入文本翻译到另一种语言（支点），然后将翻译得到的文本再次翻译回当前语言。由于每次翻译过程中均要求源语言和目标语言中文本的语义保持一致，因此可以预期最后得到的文本在语义上能跟输入文本保持一致。注意支点语言可以只采用一种语言，也可以采用多种语言。

需要指出的是，句子简化（Sentence Simplification）可以看做是一类特殊的复述生成问题，其目的是将复杂的长句改写为简单、可读性更好、易于理解的句子。在句子简化的实现上，仍然可以采用上述的各种方法，例如基于单语言的机器翻译方法【10】、基于树转换的方法【11】等。

参考资料：

[1] Knight, K., & Marcu, D. (2000, August). Statistics-based summarization-step one: Sentence compression. In AAAI/IAAI (pp. 703-710).

[2] McDonald, R. T. (2006, April). Discriminative Sentence Compression with Soft Syntactic Evidence. In EACL.

[3] Cohn, T. A., & Lapata, M. (2009). Sentence compression as tree transduction. Journal of Artificial Intelligence Research, 637-674.

[4] Clarke, J., & Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research, 399-429.

[5] Barzilay, R., & Lee, L. (2003, May). Learning to paraphrase: an unsupervised approach using multiple-sequence alignment.

[6] Fujita, A., Inui, K., & Matsumoto, Y. (2005). Exploiting lexical conceptual structure for paraphrase generation

[7] Quirk, C., Brockett, C., & Dolan, W. B. (2004, July). Monolingual Machine Translation for Paraphrase Generation

[8] Duboue, P. A., & Chu-Carroll, J. (2006, June). Answering the question you wish they had asked: The impact of paraphrasing for question answering.

[9] Max, A. (2009, August). Sub-sentential paraphrasing by contextual pivot translation.

[10] ] Wubben, S., Van Den Bosch, A., & Krahmer, E. (2012, July). Sentence simplification by monolingual machine translation.

[11] Zhu, Z., Bernhard, D., & Gurevych, I. (2010, August). A monolingual tree-based translation model for sentence simplification.