自动文摘研究进展与趋势

万小军、姚金戈

北京大学计算机科学技术研究所

个人简介： 万小军，北京大学计算机科学技术研究所教授，博士生导师，语言计算与互联网挖掘实验室负责人。研究方向为自然语言处理与文本挖掘，研究兴趣包括自动文摘与文本生成、情感分析与观点挖掘、语义计算与信息推荐等，在相关学术会议与期刊上发表高水平学术论文上百篇。担任计算语言学顶级国际期刊 Computational Linguistics 编委，TACL 常务评审委员（Standing Reviewing Committee）， 多次担任自然语言处理领域一流与重要国际会议领域主席或 SPC（包括 ACL、NAACL、IJCAI、IJCNLP 等），自主或合作研制了自动文摘开源平台 PKUSUMSUM、AI 写稿机器人 Xiaomingbot 等系统。

姚金戈，北京大学计算机科学技术研究所博士生，研究方向为自然语言处理与自动文摘。

自动文摘的目的是通过对原文本进行压缩、提炼，为用户提供简明扼要的文字描述。根据处理的文档数量，自动文摘可以分为只针对单篇文档的单文档自动摘要和针对文档集的多文档自动摘要。根据是否提供上下文环境，自动文摘可以分为与主题或查询相关的自动摘要以及普通自动摘要。根据摘要的不同应用场景，自动文摘可以分为传记摘要、观点摘要、学术文献综述生成等，这些摘要通常为满足特定的应用需求。

自动文摘可以看作是一个信息压缩过程，将输入的一篇或多篇文档压缩为一篇简短的摘要，该过程不可避免有信息损失，但是要求保留尽可能多的重要信息。自动文摘系统通常涉及到对输入文档的理解、要点的筛选，以及文摘合成这三个主要步骤。其中，文档理解可浅

可深，大多数自动文摘系统只需要进行比较浅层的文档理解，例如段落划分、句子切分、词法分析等，也有文摘系统需要依赖句法解析、语义角色标注、指代消解，甚至深层语义分析等技术。

研究现状与进展

自动文摘所采用的方法从实现上考虑可以分为抽取式摘要（extractive summarization）和生成式摘要（abstractive summarization）。抽取式方法相对比较简单，通常利用不同方法对文档结构单元（句子、段落等）进行评价，对每个结构单元赋予一定权重，然后选择最重要的结构单元组成摘要。而生成式方法通常需要利用自然语言理解技术对文本进行语法、语义分析，对信息进行融合，利用自然语言生成技术生成新的摘要句子。

目前主流自动文摘研究工作大致遵循如下技术框架：

内容表示 → 权重计算 → 内容选择 → 内容组织

首先将原始文本表示为便于后续处理的表达方式，然后由模型对不同的句法或语义单元进行重要性计算，再根据重要性权重选取一部分单元，经过内容上的组织形成最后的摘要。现有的研究工作针对不同设定和场景需求展开，为上述框架中的各个技术点提供了多种不同的设计方案。有不少相关研究也尝试在统一的框架中联合考虑其中的多个技术点。

1 内容表示与权重计算

原文档中的每个句子由多个词汇或单元构成，后续处理过程中也以词汇等元素为基本单位，对所在句子给出综合评价分数。以基于句子选取的抽取式方法为例，句子的重要性得分由其组成部分的重要性衡量。由于词汇在文档中的出现频次可以在一定程度上反映其重要性，我们可以使用每个句子中出现某词的概率作为该词的得分，通过将所有包含词的概率求和得到句子得分(Nenkova and Vanderwende, 2005; Vanderwende et al., 2007)。也有一些工作考虑更多细节，利用扩展性较强的贝叶斯话题模型，对词汇本身的话题相关性概率进行建模(Daume III and Marcu, 2006; Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2010)。

一些方法将每个句子表示为向量，维数为总词表大小。通常使用加权频数(Salton and Buckley, 1988; Erkan and Radev, 2004)作为句子向量相应维上的取值。加权频数的定义可以有多种，如信息检索中常用的词频-逆文档频率（TF-IDF）权重。也有研究工作考虑利用隐语义分析或其他矩阵分解技术，得到低维隐含语义表示并加以利用(Gong and Liu, 2001)。得到向量表示后计算两两之间的某种相似度（例如余弦相似度）。随后根据计算出的相似度构建

带权图，图中每个节点对应每个句子。在多文档摘要任务中，重要的句子可能和更多其他句子较为相似，所以可以用相似度作为节点之间的边权，通过迭代求解基于图的排序算法来得到句子的重要性得分(Erkan and Radev, 2004; Wan et al., 2007; Wan and Yang, 2008)。也有很多工作尝试捕捉每个句子中所描述的概念，例如句子中所包含的命名实体或动词。出于简化考虑，现有工作中更多将二元词（bigram）作为概念(Gillick et al., 2008; Li et al., 2013)。

另一方面，很多摘要任务已经具备一定数量的公开数据集，可用于训练有监督打分模型。例如对于抽取式摘要，我们可以将人工撰写的摘要贪心匹配原文档中的句子或概念，从而得到不同单元是否应当被选作摘要句的数据。然后对各单元人工抽取若干特征，利用回归模型(Ouyang et al., 2011; Hong and Nenkova, 2014)或排序学习模型(Shen and Li, 2011; Wang et al., 2013)进行有监督学习，得到句子或概念对应的得分。文档内容描述具有结构性，因此也有利用隐马尔科夫模型（HMM）、条件随机场（CRF）、结构化支持向量机（Structural SVM）等常见序列标注或一般结构预测模型进行抽取式摘要有监督训练的工作(Conroy, 2001; Shen et al., 2007; Sivos and Joachims, 2012)。所提取的特征包括所在位置、包含词汇、与邻句的相似度等等。对特定摘要任务一般也会引入与具体设定相关的特征，例如查询相关摘要任务中需要考虑与查询的匹配或相似程度。

2 内容选择

无论从效果评价还是从实用性的角度考虑，最终生成的摘要一般在长度上会有限制。在获取到句子或其他单元的重要性得分以后，需要考虑如何在尽可能短的长度里容纳尽可能多的重要信息，在此基础上对原文内容进行选取。

2.1 贪心选择

可以根据句子或其他单元的重要性得分进行贪心选择。选择过程中需要考虑各单元之间的相似性，尽量避免在最终的摘要中包含重复的信息。最为简单常用的去除冗余机制为最大边缘相关法(Maximal Marginal Relevance – MMR)(Carbonell and Goldstein, 1998)，即在每次选取过程中，贪心选择与查询最相关或内容最重要、同时和已选择信息重叠性最小的结果。也有一些方法直接将内容选择的重要性和多样性同时考虑在同一个概率模型框架内(Kulesza and Taskar, 2011)，基于贪心选择近似优化似然函数，取得了不错的效果。

此后有离散优化方向的研究组介入自动文摘相关研究，指出包括最大边缘相关法在内的很多贪心选择目标函数都具有次模性(Lin and Bilmes, 2010)。记内容选取目标函数为 $F(S)$，其自变量 S 为待选择单元的集合；次模函数要求对于 $\forall S \subseteq T \subseteq U \setminus u$，以及任意单元 u，

都满足如下性质：

$$F(S \cup \{u\}) - F(S) \geq F(T \cup \{u\}) - F(T).$$

这个性质被称为回报递减效应（diminishing returns），很符合贪心选择摘要内容的直觉：由于每步选择的即时最优性，每次多选入一句话，信息的增加不会比上一步更多。使用特定的贪心法近似求解次模函数优化问题，一般具备最坏情况近似比的理论保证。而实际应用中研究发现，贪心法往往已经可以求得较为理想的解。由于贪心法易于实现、运行效率高，基于次模函数优化的内容选择在近年得到了很多扩展。多种次模函数优化或部分次模函数优化问题及相应的贪心解法被提出，用于具体语句或句法单元的选取(Lin and Bilmes, 2011; Sipos et al., 2012; Dasgupta et al., 2013; Morita et al., 2013)。

2.2 全局优化

基于全局优化的内容选择方法同样以最大化摘要覆盖信息、最小化冗余等要素作为目标，同时可以在优化问题中考虑多种由任务和方法本身的性质所导出的约束条件。最为常用的形式化框架是基于0-1二值变量的整数线性规划(McDonald, 2007; Gillick and Favre, 2009)。最后求解优化问题得到的结果中如果某变量取值为1，则表示应当将该变量对应的单元选入最后的摘要中。由于整数线性规划在计算复杂性上一般为NP-难问题，此类方法的求解过程在实际应用中会表现较慢，并不适合实时性较高的应用场景。有研究工作将问题简化后使用动态规划策略设计更高效的近似解法。也有少量研究工作尝试在一部分特例下将问题转化为最小割问题快速求解(Qian and Liu, 2013)，或利用对偶分解技术将问题化为多个简单子问题尝试求得较好的近似解(Almeida and Martins, 2013)。更为通用的全局优化加速方案目前仍是一个开放问题。

3 内容组织

3.1 内容简化与整合

基于句子抽取得到的语句在表达上不够精练，需要通过语句压缩、简化、改写等技术克服这一问题。在这些技术中相对而言较为简单的语句压缩技术已经广泛被应用于摘要内容简化。现行主要做法基于句法规则(Clarke and Lapata, 2008)或篇章规则(Clarke and Lapata, 2010; Durrett et al., 2016)，例如如果某短语重要性较高需要被选择用于构成摘要，那么该短语所修饰的中心词也应当被选择，这样才能保证得到的结果符合语法。这些规则既可以直接用于后处理步骤衔接在内容选取之后进行，也可以用约束的形式施加在优化模型中，这样在求解优化问题完毕后就自然得到了符合规则的简化结果。局部规则很容易表达为变量之间的线性不等式约束，因此尤其适合在前面提到的整数线性规划框架中引入。另外，关于语句简化与改

写方面目前也有相对独立的研究，主要利用机器翻译模型进行语句串或句法树的转写(Wubben et al., 2012)。由于训练代价高以及短语结构句法分析效率和性能等诸多方面原因，目前很少看到相关模块在摘要系统中的直接整合与应用。

一些非抽取式摘要方法则重点考虑对原句信息进行融合以生成新的摘要语句。基于句法分析和对齐技术，可以从合并后的词图直接产生最后的句子(Barzilay and McKeown, 2005)，或者以约束形式将合并信息引入优化模型(Bing et al., 2015)等方式来实现。

还有部分研究者尝试通过对原文档进行语义理解，将原文档表示为深层语义形式（例如深层语义图），然后分析获得摘要的深层语义表示（例如深层语义子图），最后由摘要的深层语义表示生成摘要文本。近期的一个尝试为基于抽象意义表示（Abstract Meaning Representation, AMR）进行生成式摘要(Liu et al., 2015)。这类方法所得到的摘要句子并不是基于原文句子所得，而是利用语义分析和自然语言生成技术从语义表达直接生成而得。这类方法相对比较复杂，而且由于自然语言理解与自然语言生成本身都没有得到很好的解决，因此目前生成式摘要方法仍属于探索阶段，其性能还不尽如人意。

3.2 内容排序

关于对所选取内容的排序，相关研究尚处于较为初级的阶段。对于单文档摘要任务而言，所选取内容在原文档中的表述顺序基本可以反映这些内容之间正确的组织顺序，因此通常直接保持所选取内容在原文中的顺序。而对于多文档摘要任务，选取内容来自不同文档，所以更需要考虑内容之间的衔接性与连贯性。早期基于实体的方法(Lapata and Barzilay, 2005; Barzilay and Lapata, 2008)通过对实体描述转移的概率建模计算语句之间的连贯性。据此找到一组最优排序的问题很容易规约到复杂性为 NP-完全的旅行商问题，精确求解十分困难。因此多种近似算法已经被应用于内容排序。近年来，深度学习技术被用于语句连贯性建模与排序任务中，Li 与 Jurafsky (2016)提出基于 LSTM 的辨别式模型与生成式模型，能够取得比较理想的排序效果。未来随着篇章分析、指代消解技术的不断进步，多文档摘要中的语句排序问题也有机会随之产生更好的解决方案。

4 端到端摘要

随着深度学习技术在分布式语义、语言模型、机器翻译等任务上取得了一系列突破性成果，相关方法在文摘任务上的应用研究也受到广泛关注。基于编码器-解码器（encoder-decoder）架构的序列到序列学习模型（sequence-to-sequence learning）目前最为流行，因为可以避免繁琐的人工特征提取，也避开了重要性评估、内容选择等技术点的模块化，只需要足够的输入输出即可开始训练。但这些方法需要比传统方法规模远远更大的训练语

料，加上当前主流的神经网络框架尚不能够有效对长文档进行语义编码，因此目前的相关研究大多只能集中于语句级简化和标题生成，一般仅仅以文档首句作为输入，以一个短句作为输出(如 Rush et al., 2015; Gu et al., 2016 等)。极少数近期工作开始同时在同一个神经网络框架里考虑句子选取和摘要生成，尝试对语句层次进行编码并在此基础上引入层次化注意机制(Li et al., 2015; Cheng et al., 2016)，但效果尚未能明显改善传统方法已经能够取得的性能。

展望

自动文摘是自然语言处理领域的一个重要研究方向，近 60 年持续性的研究已经在部分自动文摘任务上取得了明显进展，但仍需突破很多关键技术，才能提高其应用价值、扩大其应用范围。

展望未来，以下研究方向或问题值得关注：

多语言自动文摘资源建设：目前的自动文摘资源总体上偏少，无论是数据还是工具与系统。一方面会影响评测结果的准确性，另一方面也无法为有监督学习方法尤其是深度学习方法提供充足的训练数据。业界需要投入更多的人力物力来建设多语言自动文摘资源。

自动文摘评价方法的完善：目前的自动文摘评价方法需要进一步完善，尤其是自动评价方法。基于词汇重叠程度的 ROUGE 等评价方法虽然被广泛采用，但质疑声不断。业界需要提出更加合理的自动评价准则，综合考虑摘要的多种性质，这将极大推动业界对自动文摘的研究。

基于自然语言生成的自动文摘：生成式摘要方法更符合人类撰写摘要的习惯，但自然语言生成技术的复杂性和不成熟阻碍了生成式摘要方法的研究进展。深度学习技术在自然语言生成问题上的逐步应用给生成式摘要带来了希望和机遇，未来几年将会有越来越多的研究者基于深度学习技术从事生成式摘要方法的研究，也有望取得重要进展。

篇章信息和语义信息的有效利用：现有方法利用的信息主要基于由统计频数或出现位置所反映的重要性度量，一般比较表层，而忽视了对文档篇章信息与语义信息的利用。文档本身的语义表达具备很强的结构性，各语义单元之间存在紧密联系，这一点在目前提出的结构预测模型中也几乎没有考虑。另一方面，应尽可能保证最后抽取或生成的摘要在描述上前后一致、表达连贯。因此，对文档篇章与语义信息的有效利用将有可能大大改善自动文摘系统的性能。

综述自动生成：综述自动生成是一类特殊的自动文摘任务，具有广泛的应用价值，可帮助自动撰写新闻事件深度报道、学术文献综述、舆情报告等。与传统自动文摘任务不同，综

述一般较长，可以长达数千字，牵涉到篇章的整体逻辑性与局部连贯性，因此更具有挑战性。目前业界仅仅对学术文献自动综述进行了简单了尝试，效果差强人意，未来几年期待业界研究者在更多综述自动生成任务上进行有益的尝试，并在特殊应用场景下实现风格相对固定的综述文章自动撰写。

面向复杂问题回答的自动摘要：基于关键词检索的搜索引擎正在逐步向基于自然语言检索的问答引擎过渡。而对于很多种类的问题，并不适合使用简单的一两个短语作答。比如搜索引擎用户时常需要进行对定义（“是什么”）、原因（“为什么”）、步骤（“怎么做”）、观点（“怎么样”）等方面的查询。与只需少量简单实体作答的事实型问题相对，这一类问题往往被称为非事实型问题或复杂问题。相对完整地回答非事实型问题需要对单个文档甚至多个相关文档中的部分内容进行提取、聚合与总结。由于非事实型问答固有的困难性，相关研究在学术圈进展缓慢，期待未来有更多的研究者敢于迎接此项挑战。

除了上述研究方向与问题之外，未来自动文摘将会越来越多地与其他技术相结合，面向全新的应用需求，形成更具特色的自动文摘任务，该领域的研究也将更加多样化。

最后，我们有理由相信，随着语义分析、篇章理解、深度学习等技术的快速发展，自动文摘这一重要且有挑战性的自然语言处理问题在可预见的未来能够取得显著的研究进展，并且更多地应用于互联网产品与服务，从而体现自身的价值。

参考文献

- Almeida, M. B., & Martins, A. F. (2013). Fast and Robust Compressive Summarization with Dual Decomposition and Multi-Task Learning. In ACL.
- Bing L., Li P., Liao Y., Lam W., Guo W., & Passonneau R. J. (2015). Abstractive Multi-Document Summarization via Phrase Selection and Merging. In ACL.
- Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. Computational Linguistics, 34(1), 1-34.
- Barzilay, R., & McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3), 297-328.
- Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In SIGIR.
- Celikyilmaz, A., & Hakkani-Tur, D. (2010). A hybrid hierarchical model for multi-document summarization. In ACL.
- Cheng, J., & Lapata, M. (2016). Neural Summarization by Extracting Sentences and Words. In ACL.
- Clarke, J., & Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research, 399-429.
- Clarke, J., & Lapata, M. (2010). Discourse constraints for document compression. Computational Linguistics, 36(3), 411-441.
- Conroy, J. M., & O'leary, D. P. (2001). Text summarization via hidden markov models. In SIGIR.

- Dasgupta, A., Kumar, R., & Ravi, S. (2013). Summarization Through Submodularity and Dispersion. In ACL.
- Daumé III, H., & Marcu, D. (2006). Bayesian query-focused summarization. In ACL.
- Durrett, G., Berg-Kirkpatrick, T., & Klein, D. (2016). Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints. In ACL.
- Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457-479.
- Gillick, D., Favre, B., & Hakkani-Tur, D. (2008). The ICSI summarization system at TAC 2008. In Proceedings of the Text Understanding Conference.
- Gillick, D., & Favre, B. (2009). A scalable global model for summarization. In Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing (pp. 10-18). Association for Computational Linguistics.
- Gu, J., Lu, Z., Li, H., & Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. In ACL.
- Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In ACL.
- Hong, K., & Nenkova, A. (2014). Improving the Estimation of Word Importance for News Multi-Document Summarization. In EACL.
- Kulesza, A., & Taskar, B. (2011). Learning determinantal point processes. In UAI.
- Lapata, M., & Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In IJCAI.
- Li, C., Qian, X., & Liu, Y. (2013). Using Supervised Bigram-based ILP for Extractive Summarization. In ACL.
- Li, J., & Jurafsky, D. (2016). Neural Net Models for Open-Domain Discourse Coherence. arXiv, https://arxiv.org/abs/1606.01545v1.
- Li, J., Luong, M. T., & Jurafsky, D. (2015). A hierarchical neural autoencoder for paragraphs and documents. In ACL.
- Lin, H., & Bilmes, J. (2010). Multi-document summarization via budgeted maximization of submodular functions. In HLT-NAACL.
- Lin, H., & Bilmes, J. (2011). A class of submodular functions for document summarization. In HLT-NAACL.
- Lin, C. Y., & Hovy, E. (2002). From single to multi-document summarization: A prototype system and its evaluation. In ACL.
- Liu, F., Flanigan, J., Thomson, S., Sadeh, N., & Smith, N. A. (2015). Toward Abstractive Summarization Using Semantic Representations. In NAACL.
- McDonald, R. (2007). A study of global inference algorithms in multi-document summarization (pp. 557-564). Springer Berlin Heidelberg.
- Morita, H., Sasano, R., Takamura, H., & Okumura, M. (2013). Subtree Extractive Summarization via Submodular Maximization. In ACL.
- Nenkova, A., & Vanderwende, L. (2005). The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101.
- Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. Information Processing & Management, 47(2), 227-237.
- Qian, X., & Liu, Y. (2013). Fast Joint Compression and Summarization via Graph Cuts. In EMNLP.

- Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In EMNLP.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.
- Sipos, R., Shivaswamy, P., & Joachims, T. (2012). Large-margin learning of submodular summarization models. In EACL.
- Shen, C., & Li, T. (2011). Learning to rank for query-focused multi-document summarization. In ICDM.
- Shen, D., Sun, J. T., Li, H., Yang, Q., & Chen, Z. (2007). Document Summarization Using Conditional Random Fields. In IJCAI.
- Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management, 43(6), 1606-1618.
- Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In SIGIR.
- Wan, X., Yang, J., & Xiao, J. (2007). Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In IJCAI.
- Wang, L., Raghavan, H., Castelli, V., Florian, R., & Cardie, C. (2013). A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization. In ACL.
- Wubben, S., Van Den Bosch, A., & Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In ACL.