**文本真实性的分类**

在典型的NLP应用中，信息的真实性对句子和文章的理解起着非常关键的作用。

文本真实性的研究中，有两个重要的概念——推测和否定。推测是一种确定和否定之间的不确定表达式，而否定是一种颠倒命题真值的语法范畴。

国内外现状：

推测和否定提取在近些年受到了广泛的关注，因为它能较好的区分事实和反事实、不确定的信息之前的差别。

通常，推测和否定提取涉及两个典型的子任务：提示识别（cue identification）和范围检测（scope detect）。这里，提示是指具有推测和否定意义的词语或者短语（例如suspect、guess、deny、not），范围是句子中提示（提示词）所管理的文章片段。

Tang 等人在2011在提示识别方便取得了非常大的成功（Tang B, Wang X, Wang X, et al. A cascade method for detecting hedges and their scope in natural language text[C]// 2010:13-17.）。相比较，推测和否定的范围检索依然是一个挑战，因为它本身的困难和那些检索范围的上文中错误。

关于有关范围检索的研究，有通过基于启发式规则的方式(Lilja Øvrelid, Erik Velldal, and Stephan Oepen. 2010. Syntactic Scope Resolution in Uncertainty Analy-sis. In Proceedings of the 23rd International Con-ference on Computational Linguistics (Coling 2010), Beijing, August 2010, pages 1379-1387.)，有使用机器学习的方式（Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2013. Tree Kernel-based Negation and Speculation Scope Detection with Structured Syntactic Parse Features. In Proceedings of the 2013 Conference on Empiri-cal Methods in Natural Language Processing (EMNLP 2013), Seattle, Washington, USA, 2013, pages 968-976.），还有使用规则整合的经验模型（Velldal et al., 2012）。在2016年，Qian等人（Qian Z, Li P, Zhu Q, et al. Speculation and Negation Scope Detection via Convolutional Neural Networks[C]// EMNLP 2016: Conference on Empirical Methods in Natural Language Processing. 2016.）使用了基于CNN的方式作用在的范围检测方面。他们的研究使用的CNN模型可以从语法路径中在提示词和从候选区和依赖分析树获得的候选词之间提取多种不同的有用的特征。他们的模型也可以受益于神经网络从语法路径或更复杂的句法树中提取有用信息的能力。

**立场分类（姿态分类）**

概念解释：

立场分类是从文本中推断出，作者是否赞成给定目标，或者反对它，还是对它持中立态度。这项任务对人来说都是很困难的，它与论证挖掘、主观性分析和情绪分类都有关系。一般的情绪分类，可以公式化为确定一段文本是正面的、负面的还是中性的三分类问题。然而，在立场分类中，系统必须检测对给定（预选选择的）重要目标的好感度（favorability）。在这个意义上，立场分析更类似于目标依赖的情感分类（Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification In Proceedings of ACL, pages 151–160.），主要区别在于，文本中可能没有明确的提及立场的目标，或者没有意见的目标。

国内外现状：

以前对立场分析的研究工作集中在国会辩论、公司内部讨论和在线论坛的辩论的文本立场分析。Somasundaran和Wiebe（Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 116–124.）使用基于模态动词和情绪的特征开发了立场分类的基线。Anand 在2011年利用基于词典和基于依赖的特征增强了n-garm特征。立场分类刚被充当为集体分类，引用结构或者反驳链接被用作额外的信息来创立争论帖子中的支持和不支持的模型并推断出他们的标签。（Pranav Anand, MarilynWalker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, pages 1–9.）。在2010年Murakami和Raymond（Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In Proceedings of COLING, pages 869–875.）使用最大切割的方法用于聚合多个帖子中的立场来推断出用户对目标的立场。Sridhar使用概率软逻辑（PSL）集体分类用户的立场和帖子中的立场（Dhanya Sridhar, James Foulds, Bert Huang,Lise Getoor, and MarilynWalker. 2015. Joint models of disagreement and stance in online debate. In Proceedings of ACL, pages 116–125.）。

分类在线微博社交媒体中表达的立场是观念挖掘中一个新问题。Javid Ebrahimi在2016年提出一个种概率分析方法，用于推特中建立立场分类，它创造了立场、立场的目标和推特的情感。不仅仅简单的将情感或者目标变量结合到特征空间，他们还使用一个新颖的公式来结合情感状态输入变量中的三项交互和目标状态输入变量之间的三项交互。他们提出的规范直观目的为立场分类从目标特征中区分出情感特征。

Ferreira 讨论了研究情绪和立场之间的相关性以及前者如何帮助检测后者（Ferreira W, Vlachos A. Emergent: a novel data-set for stance classification[C]// The, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.）。他们使用无向图形模型来解决目标、立场本身和文本中情感之间的相互作用的建模。

他们使用一个单一的分类器对多个目标的进行立场分类，为每个目标使用单独的分类器。他们演示了特定模型规范和共享正则化可以如何改善多个目标之间的姿态分类。

**主题分类**：

主题分类现在面临着文档之间巨大的变化性的困难，因为近些年自由形式无结构文本的几何倍数的增长。本文将变化性视为文档之间的差异性，目的是为了降低文本之间主题的变化性为了获得更好的文本分类。有许多导致文本之间主题的原因，比如说不同的人所使用的不同的语言用法。不同的人会用不同的词来表达他们的观点，因此同一主题的文章会很不相同因为主题内部多变性的问题。

传统解决内部主题变化的问题方法——高维的TF-IDF向量。TF-IDF没有揭示术语之间的相似性信息，这将增加使用不同词文章之间的差异。应用无监督技术的潜在语意分析方法，典型的方法就是潜在语意索引（LSI），而由LSI估计的特征是原始特征的线性组合（Quan Wang, Jun Xu, Hang Li, and Nick Craswell. 2013.Regularized latent semantic indexing: A new approach to large-scale topic modeling. ACM Transactions on Information Systems (TOIS), 31(1):5.）。同时还有特别流行的潜在的狄里克雷的分配算法通过生成概率模型来代表文章（Mohamed Morchid, Richard Dufour, and Georges Linares. 2014. A lda-based topic classification approach from highly imperfect automatic transcriptions. LREC14.）。

由于以前的工作没有直接解决讨论组中主题分歧的解析，Amanda Chow和Jenny Hong提出的主题模型比以前更加丰富，包括频率逆文档频率（TF-IDF），潜在狄利克雷分析（LDA）和监督LDA（sLDA）。(Topical Classification and Divergence on Reddit)这些都是流行和建立的主题模型。但是这些主题模型使用的是词的一元模型或者N-gram特征，但不是词向量。Amanda Chow和Jenny Hong在他们的项目中使用神经网络的词向量。

关于主题分类的研究包括两个任务。首先，训练神经分类器来给预测所给的评论文本的类。然后我们应用分类器讨论某个“深度”的评论集合然后找到特定的讨论（thread ）。

他们研究主要处理主题内部变化性问题。这项工作的主要目的是找到一个具有最小化的内部主题变化性的新的特征空间。构建客观标准用于优化。数学上，他们利用训练集的主题标签信息来构造一个加权矩阵，然后对所有内在主题文章的差异求和。然后，通过是用解决最优解问题的基于KPCA的有效的算法，产生强大的特征空间来最小化内部主题的变化性。最后，完成基线特征的变化性的归一操作。