引文分析综述

# 一．语义引文上下文识别 （后续的基础）

使用论证挖掘算法来建立引文上下文的长度，以自动检测引文周围的潜在上下文。探究引文窗口大小，即引文上下文的分布位置，

1. Nanba和Okumura（1999）使用引用区域来引用相同的概念。他们将引用区域定义为在科学论文中围绕给定参考文献的位置传递的句子的连续性，并与之相关联。他们的算法开始于将包含目标参考的句子作为引用区域中的第一个成员句子。然后，他们使用一组提示词和手工制作规则来确定是否应该在引用区域添加句点。在（Nanba等人，2000）中，他们使用它们的算法来改进引文类型分类和自动测量生成。
2. Kataria(Kataria, Mitra, Caragea, & Giles, 2011)认为上下文窗口与引文与被引文之间的主题相似度相关，关系越大，窗口越大。
3. 有些认为论文结构有关。(Liakata, Saha, Dobnik, Batchelor, & Rebholzschuhmann, 2012) 创建了一个 SVM和CRF的工具来训练和比较256篇论文，根据CoreSCs把文章分成11个类别Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result, and Conclusion 。
4. Athar (Athar & Teufel, 2012a)使用支持向量机（SVM）开发了一个分类器，特征：首字母缩略词和词汇替代（即目标论文的所有引用句子中最常用的大写词，作者有时使用除首字母缩略词，代词和工作名词之外的词汇来替代引用）。

# 二．引文情感分析

## 2.1 情感分析

Sebastiani（2002）研究的结论是应用于情感分析的最佳分类器是超向量机。

为了对引用的情感极性进行分类，可以使用不同程度的有效性的不同特征集。由于语料库的多样性，不同研究的结果是不可比的，但是我们可以开始使用作者的结果来评估引文极性分类中不同特征的影响。也应考虑Hedging，因为它被用来伪装作者对被引作品的消极处置，但尚未被用于这项任务。

Jochim和Schütze(Jochim & Schutze, 2012)收集并分类了以前工作中使用的不同类型的特征，并评估了它们对不同方面分类的影响。

Athar（2011）使用POS标签的字级语言特征来实现极性分类。Abu-Jbara et al。 （2013年）使用否定，主观性，猜测和其他类似提示的列表进行BoW分析来检测极性。

Teufel等（2009），Dong和Scḧafer（2011）以及Jochim和Schütze（2012）在论文的句子，段落或部分中显示了引文功能与极性之间的联系。 Teufel等（2006）强调了引文位于其功能和极性附近的论文部分之间的关​​系。

对于极性引文分类，Athar（2014）使用了一个由852篇论文组成的语料库，引用了前20名目标论文。 Athar通过使用1-3克，依赖特征和基于窗口的否定的朴素贝叶斯分类器（宏F = 0.471，微F = 0.755）获得了SVM分类器（宏F = 0.808，微F = 0.764）的最佳结果。依赖性结构是三元组，或关系（头，依赖），捕获词之间的长距离关系。基于窗口的否定认为任何否定项的k字窗口（k = 1-15）内的所有单词都是后缀，以区别于非极性版本。极性分类和引用功能的不同方案类别如表1所示。

Athar（2014）报道引用阴性极少的引用次数较少。其中一些负面提及不能被发现，因为作者往往通过使用套期保值来隐瞒引用论文的负面处置。有效的对冲检测是一项必要的任务，需要考虑语言线索，Hyland（1996）详述。应使用技术来解决更多丰富的中性引文，从而使分类结果偏斜。

## 2.2 情感加上下文

Athar和Teufel观察到，在评估引文情绪时考虑到上下文将负引文数量增加了3倍。Athar提出了两种利用上下文的方法。在第一种方法中，他们(Athar & Teufel, 2012a)将引用句和固定上下文（引用句子周围的四个声明的窗口）视为单个句子。它们从合并的文本中提取特征，并对其分类器进行了类似于他们在2011年的论文中的描述。在第二种方法中(Athar & Teufel, 2012b)，它们使用四类注释方案。在引文周围的四个句子的窗口中的每个句子被标记为正，负，中立或排除（与引用的工作无关）。实验令人惊讶地给出了负面结果，表明在不考虑上下文的情况下分类情绪可以取得更好的效果。他们归因于它们的训练数据的小尺寸和包括上下文文本引入数据的噪声。

## 2.3 半监督

### 2.3.1 深度学习

(Jochim & Schutze, 2014)将引用分类框架作为域适应任务，并利用其他领域中可用的丰富标签数据。然后，为了避免特定科学领域的过度设计特定引用特征，研究了一种深入学习的神经网络方法（MSDA）(Chen, Xu, Weinberger, & Sha, 2012)，已经显示出使用单值和双字组特征在各个领域中得到广泛的推广。

对于分类，我们使用Chen等人的边缘化堆叠去噪自动编码器（mSDA） （2012）6加一个线性SVM。 mSDA采用降噪概念 - 引入噪声，使自动编码器更加健壮 - 来自Vincent等。 （2008），但是以封闭形式进行优化，从而避免迭代输入向量以随机引入噪声。 这样做的结果是更快的运行时间和目前MDSD上的最先进的性能，这使得它成为我们的域适配任务的不错选择。 mSDA实现带有LIBSVM，我们用LIBLINEAR（Fan et al。，2008）替换了更快的运行时间，精度没有降低。 具有默认设置的LIB-LINEAR也可作为我们的基线。

去噪自动编码器（SDA）是针对部分和随机损坏重建输入数据的单层神经网络。这些消音器可以堆叠成深入学习的架构。然后将其中间层的输出用作SVM的输入特征（Lee et al，2009）。 Glorot等人（2011）表明，使用SDA学习的特征与线性SVM分类器结合起来，可以在不同产品领域的情绪分析的基准任务上获得记录性能（Blitzer等，2006）。尽管SDA具有显着且有前途的成果，但由于其高的计算成本受到限制。与竞争算法相比，它们的训练速度要慢一些（Blitzer等，2006; Chen等，2011a; Xue et al。，2008），主要是因为它们依赖于迭代和数值优化来学习模型参数。挑战进一步加剧了输入数据的维度和计算密集型模型选择过程调整超参数的需求。因此，即使高度优化的实施（Bergstra等，2010）也许需要几个小时（甚至几天）的培训时间。

Jochim用SDA的变体解决了这个挑战。所提出的方法称为边缘化的叠加去噪自动编码器（mSDA），采用SDA的贪婪逐层训练。然而，最重要的不同之处在于我们使用线性去噪器作为基本构建块。在这种情况下随机特征的出错可能被边缘化。从概念上讲，这相当于使用无数大量损坏的输入数据对模型进行训练。对于常规SDA（通常依赖于随机梯度下降）而需要扫描所有训练数据的传统SDA，这种规模的拟合模型是不可能的。

我们的贡献总结如下：i）我们通过证明线性消除器可以用作学习特征表征的基础，来深入学习。 ii）我们表明线性度可以显着简化参数估计 - 我们的方法得到最佳参数的闭式解决方案。 iii）我们严格地对已建立的域适应基准数据集进行评估，并与几种竞争最先进的算法进行比较。我们显示，mSDA的分类性能与我们的基准数据集中的SDA的分类性能相匹配，同时在训练时间内实现了巨大的加速（通过mSDA将训练从最长达2天的SDA减少到几分钟）。

在深度学习文献中已经开发了各种形式的编码器（Rumelhart等，1986; Baldi＆Hornik，1989; Kavukcuoglu等，2009; Lee等，2009; Vincent et al。，2008 ; Rifai et al。，2011）。例如，Lee et al.（2009）表明，由卷积神经网络（CNN）的全部或部分层产生的隐藏表示形成了SVM分类的极好特征。 CNN的预处理通过增加对噪声和标签不变变换的鲁棒性来改善泛化。Glorot等人（2011）在文件情绪分析中成功应用SDA来提取域适应的特征。作者训练SDA重建源和目标数据的并集上的输入向量（忽略标签）。在所得到的特征表示h（x）上训练的分类器（例如，线性SVM）从直接传播到x上的传播明显好于源到目标。与CNN类似，SDA还将相关的输入尺寸相结合，因为它们从未破坏的特征重构去除的特征值。显示SDA能够解开隐藏的因素，解释输入数据的变化，并根据与这些因素的相关性自动对特征进行分组（Glorot等，2011）。这有助于跨域传输，因为这些通用概念对于域特定的词汇是不变的。

我们通过以无监督的方式在源和目标数据集的并集上首先学习特征来应用mSDA进行域适配。 （Glo- rot et al。，2011）报道的一个观察结果是，如果有多个领域可用，与仅在源和目标上的预培训相比，在所有领域内共享无监督的SDA预训练是有益的。我们观察到与我们的方法类似的趋势。第5节报告的结果基于从所有可用域的数据中学习的特征。一旦mSDA被训练，挤压后的所有层的输出tanh（Wtht-1）与原始特征h0相结合，形成新的表示。所有输入都转换为新功能空间。然后对转换的源输入进行线性支持向量机（SVM）（Chang＆Lin，2011）的训练，并在目标域上进行测试。 mSDA中有两个元参数：腐败概率p和层数l。在我们的实验中，两者都对来自源域的标记数据进行了5倍的交叉验证。由于mSDA培训几乎是即时的，这种网格搜索几乎完全由SVM训练时间主导。

mSDA这个方法已经被广泛应用于情感分析提取特征的方法，但是引文方面目前就发现一个(Jochim & Schutze, 2014)，即将带标签的亚马逊商品评论通过域适应应用于无标签的引文情感分析。

### 2.3.2 word2vec

提出的方法通过对从Antholgy集合（ACL-Embeddings）学习的单词嵌入进行平均来构造句子向量（sent2vec）。还研究了极性特定词嵌入（PS-Embeddings），用于分类正和负引文。句子向量形成了一个特征空间，被引用的引用句子映射到该特征空间。将这些特征输入分类器（支持向量机）进行监督分类。

# 三．引文动机分析

## 3.1 动机

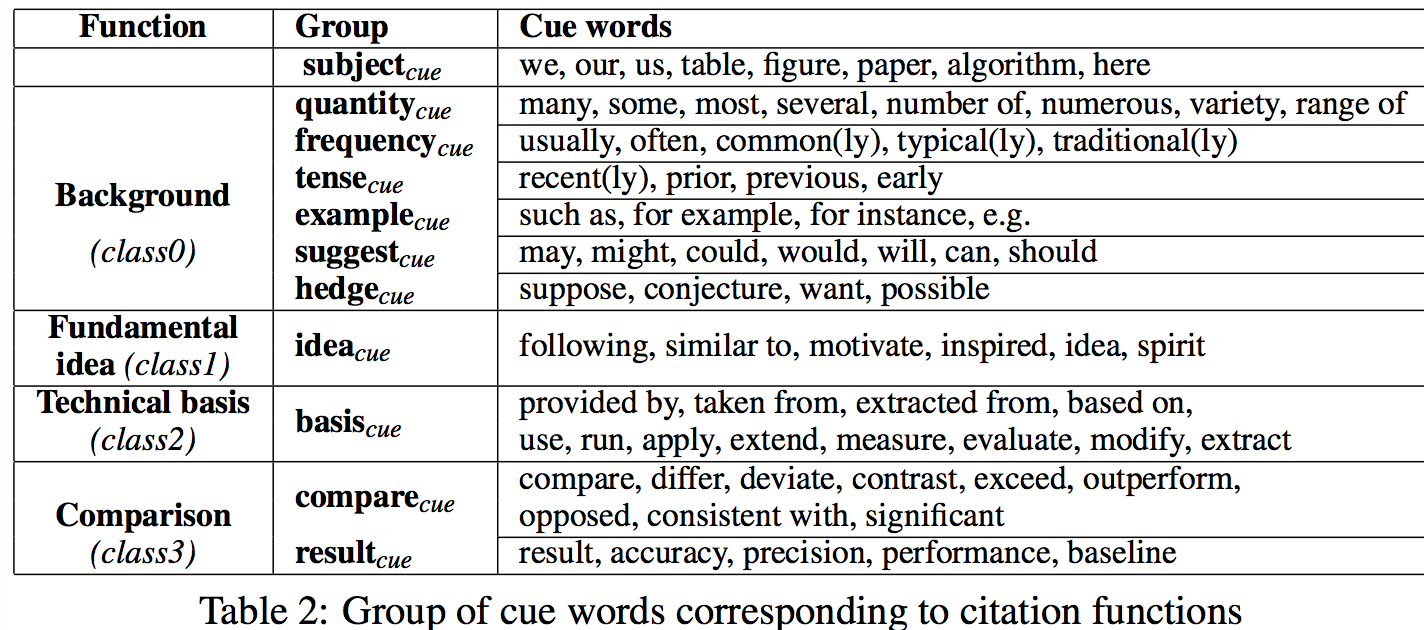
源于Garfield (Garfield, 1965)提出15种引文动机。后面Garzone又提出35种(Garzone & Mercer, 2000)。他们是用统计学的方法得到类别。随着机器学习的普及，研究者开始通过机器学习的方法进行动机分类。

Teufel等(Teufel, Siddharthan, & Tidhar, 2006)在引用函数分类工作中，采用了Spiegel-Rosing分类法的12个类别。他们训练了一个SVM分类器，并用它来标记每个引用句子。此外，他们把十二个类别映射到四个顶级类别，即：弱点，对比度（4类），正（6类）和中性。使用提示词，词性（POS）标签和不同长度的n-gram来训练WEKA7机器学习者IBk算法用于功能分类，并且获得了0.57的macro F。在同一个实验中，情感极性分类结果呈现0.71的macro F。

语料库 <http://www.cl.cam.ac.uk/~sht25/CFC.html> 只有动机

方案 <http://www.cl.cam.ac.uk/~sht25/Project_Index/Citraz_Index.html>

Dong和Scḧafer(Dong & Schafer, 2011)测试了文本，物理（位置，参考文献数量等），语料库是随机选择在2007年和2008年的计算语言学选集学会ACL会议的120篇论文。在Weka中提供的以下5种机器学习算法被用作基本分类器：BayesNet，NaiveBayes，SMO（多核选择为支持向量核），J48和RandomForest，结果为引文功能分类的F1宏观值为0.66。他们在大约170个训练实例中使用了一个半监督的（Ensemble-style Self-training）自动数据注释和集合式的自我训练算法。上下文多达三句话。他们的分类标准如下：



语料库 <https://aclbib.opendfki.de/repos/trunk/citation_classification_dataset/> 动机和情感

(Jochim & Schutze, 2012)的工作使用独立于自动分类开发的引用标准分类方案，因此不受任何特定引用应用的约束。第二，提供了一个手动注释的语料库作为进一步引用分类研究的基准。第三，引入引用分类的新功能，并将其与以前提出的引用特征进行实验比较，表明这些新特征提高了分类精度。

使用Moravcsik和Murugesan（1975）的模型从四个方面来确定引用的动机。

语料库 <http://www.ims.uni-stuttgart.de/institut/mitarbeiter/jochimcs/citation-classification> 动机

(Meyers, 2013)分为两个功能：证实Corroborate和对比。证实描述了两种遵循相同方法的作品，而对比则描述了两种不同的方法或观点。作者使用词汇词典和手工编写的规则来创建基于树模型的话语的算法。结果是获得了291个引文关系的216个正确答案。该算法适用于20篇PubMed9科学文章，对比召回率为67％，证实为83％。这些结果需要在较大的语料库中被证明。

Iorio, A., Di, Nuzzolese, A. G., and Peroni, S. 2013. Towards the Automatic Identification of the Nature of Citations. 通过模式匹配实现了引文提取算法。他们使用CiTalO检测到极性，CiTalO是使用自然语言本体学习技术，情感分析，词义消歧（与SVM）和本体映射相结合开发的软件。该语料库是关于Web标记技术的“Balisage Proceedings 10”（2011）第七卷中发表的18篇论文。最高召回率只有0.491，结果需要改进。 CiTalO实现CiTO11本体的属性。

## 3.2 动机加上下文

(Abujbara, Ezra, & Radev, 2013)使用了训练有素的分类模型和具有线性内核和ACL Anthology Network语料库的SVM分类器。他们选出了30篇论文，3500条引文。对于开发数据，他们使用序列标签来识别引用语境的方法，然后使用监督方法来确定给定引用的特征。采用一个包含六个类别的方案。目的（功能）分类结果的宏观F为58％。对于自动分类，随着更多的培训数据的提供，精度可能会增加。

他们将上下文识别视为序列标签问题，并创建了另一个300句的数据集，引用了AAN的五篇论文来确定最佳的引文上下文窗口。他们发现上下文几乎总是落在窗口之内四句话。窗口包括引用语句，引用句子前一句，引用句后两句。使用一级链式结构的条件随机字段（CRF）进行序列标记。

对于动机分析，他们使用一种监督方法，通过对从一组标签化的上下文中提取的许多词汇和结构特征对分类模型进行培训。

对于情感分析，他们采用了监督的方法。使用和以上相同特征来训练分类模型。由于数据的高度偏倚（超过一半的引用是中性的），我们使用两个设置进行二进制分类。在第一次设置中，引用被归类为极化（主观）或（中性）目标。在第二次，主观引用被归类为阳性或阴性。我们发现这种方法给出比使用3路分类器更直观的结果。

结果的影响是训练数据的大小。随着更多的培训数据的提供，准确性会提高。

## 3.3 总结

一般来说，使用SVM算法获得引文功能或目的分类的最佳结果，F1值高达73.1，这取决于所选择的特征。  
 引文分类中的主要未解决问题是缺乏功能的标准分类方案和开发具有足够特征的公共语料库，用于执行引文分类，在精确度，召回和F-度量方面产生良好结果。 目前，获得这些结果的低值，因为在大多数方案中，他们的训练注释语料库很小，并且没有为分类算法提供足够的信息来提供足够的信息。

# 四．引文重要性判定

# 五．我的想法

## 5.1 动机和情感交叉比对

## 5.2 神经网络

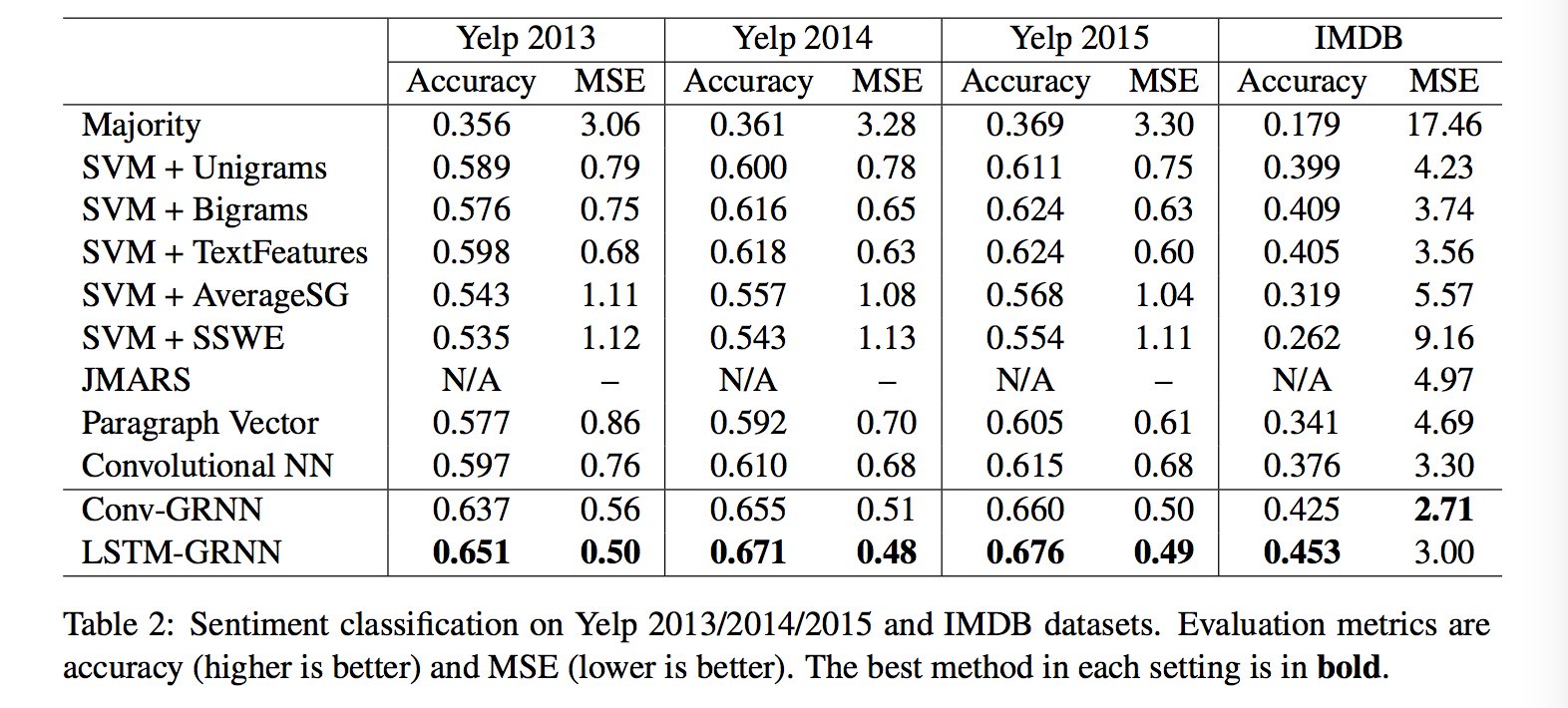
### 5.2.1 使用神经网络对文档级别的文本进行情感分类

(Tang, Qin, & Liu, 2015)引入了一个神经网络模型，以统一的自下而上的方式学习基于向量的文档表示，用于分类文档级别的情感。该模型首先用卷积神经网络（CNN）或长时间记忆（LSTM）学习句子表示。之后，句子的关系和语义被自适应地编码在具有封闭递归神经网络（gated recurrent neural network）的文档中。 我们从IMDB（大型电影评论数据集）和Yelp数据集挑战的四个大型审查数据集执行文件级别情绪分类。

Tang他们比较了多种神经网络，诸如段落向量paragraph vector（Le和Mikolov，2014），卷积神经网络CNN和基线（如基于特征的SVM）（Pang等，2002），推荐算法JMARS等表明（1）我们的神经模型显示出比几种最先进的算法更好的性能; （2）封闭递归神经网络在情感分类的文献建模中大大优于标准的递归神经网络（recurrent neural network）。

该方法具体通过两个步骤模拟文档表示。在第一步，它使用卷积神经网络（CNN）或长时间记忆（LSTM）从词表示（representations）中产生句子表示。第二步，封闭递归神经网络被用来自适应地编码在文档表示中句子的语义及其内在关系。这些表示自然地被用作对每个文档的情感标签进行分类的特征。整个模型训练端对端的随机梯度下降，其中损失函数是监督情绪分类的交叉熵误差。（The entire model is trained end-to-end with stochastic gradient descent, where the loss function is the cross-entropy error of supervised sentiment classification。）

该模型基于组成原理（Frege，1892）来记录语义，其中指出，较长表达式（例如句子或文档）的含义来自其组成成分的含义和用于组合它们的规则。由于一个文件是由一个句子列表组成，每个句子都是由一个单词组成的，所以该方法将文档表示分为两个阶段。它首先从具有句子组合的单词表示产生连续的句子向量（2.1节）。之后，句子向量被视为文献构成的输入，以获得文件表示（第2.2节）。然后将文档表示用作文档级别的分类功能（第2.3节）。



### 5.2.2 Semi-supervised sequence tagging with bidirectional language models

由于简单性和有效性，预处理的词嵌入（word embedding）在NLP系统中已经普遍存在。许多以前的研究表明，它们捕获有用的语义和句法信息（Mikolov等，2013; Pennington等，2014），并将它们包括在NLP系统中已被证明对于各种下游任务非常有帮助（Collobert et 2011）。然而，在许多NLP任务中，重要的不仅仅是表达单词的含义，而且还表示上下文中的单词。例如，在“A中央银行发言人”和“中非共和国”两个词中，“中央”一词被用作组织和地点的一部分。因此，现有技术的序列标签模型通常包括双向重写，

目前的神经网络（RNN），在进行令牌特定预测之前将令牌序列编码为上下文敏感表示（Yang等，2017; Ma和Hovy，2016; Lample等，2016; Hashimoto等，2016）。虽然令牌表示是通过预先训练的嵌入进行初始化的，但双向RNN的参数通常仅在标记数据上学习。以前的工作已经探索了通过其他任务的补充标签数据共同学习双向RNN的方法（例如，Søgaard和Goldberg，2016; Yang等，2017）。

# 六．参考文献

Abujbara, A., Ezra, J., & Radev, D. R. (2013). *Purpose and Polarity of Citation: Towards NLP-based Bibliometrics.* Paper presented at the north american chapter of the association for computational linguistics.

Athar, A., & Teufel, S. (2012a). *Context-enhanced citation sentiment detection.* Paper presented at the north american chapter of the association for computational linguistics.

Athar, A., & Teufel, S. (2012b). *Detection of implicit citations for sentiment detection*.

Chen, M., Xu, Z., Weinberger, K. Q., & Sha, F. (2012). *Marginalized Denoising Autoencoders for Domain Adaptation.* Paper presented at the international conference on machine learning.

Dong, C., & Schafer, U. (2011). *Ensemble-style Self-training on Citation Classification.* Paper presented at the international joint conference on natural language processing.

Garfield. (1965). Can Citation Indexing Be Automated?

Garzone, M., & Mercer, R. E. (2000). Towards an Automated Citation Classifier. In H. J. Hamilton (Ed.), *Advances in Artificial Intelligence: 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000 Montéal, Quebec, Canada, May 14–17, 2000 Proceedings* (pp. 337-346). Berlin, Heidelberg: Springer Berlin Heidelberg.

Jochim, C., & Schutze, H. (2012). *Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme.* Paper presented at the international conference on computational linguistics.

Jochim, C., & Schutze, H. (2014). *Improving Citation Polarity Classification with Product Reviews.* Paper presented at the meeting of the association for computational linguistics.

Kataria, S., Mitra, P., Caragea, C., & Giles, C. L. (2011). *Context sensitive topic models for author influence in document networks.* Paper presented at the international joint conference on artificial intelligence.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholzschuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics, 28*(7), 991-1000.

Meyers, A. (2013). *Contrasting and Corroborating Citations in Journal Articles.* Paper presented at the recent advances in natural language processing.

Tang, D., Qin, B., & Liu, T. (2015). *Document Modeling with Gated Recurrent Neural Network for Sentiment Classification.* Paper presented at the empirical methods in natural language processing.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). *Automatic classification of citation function*. Paper presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.